

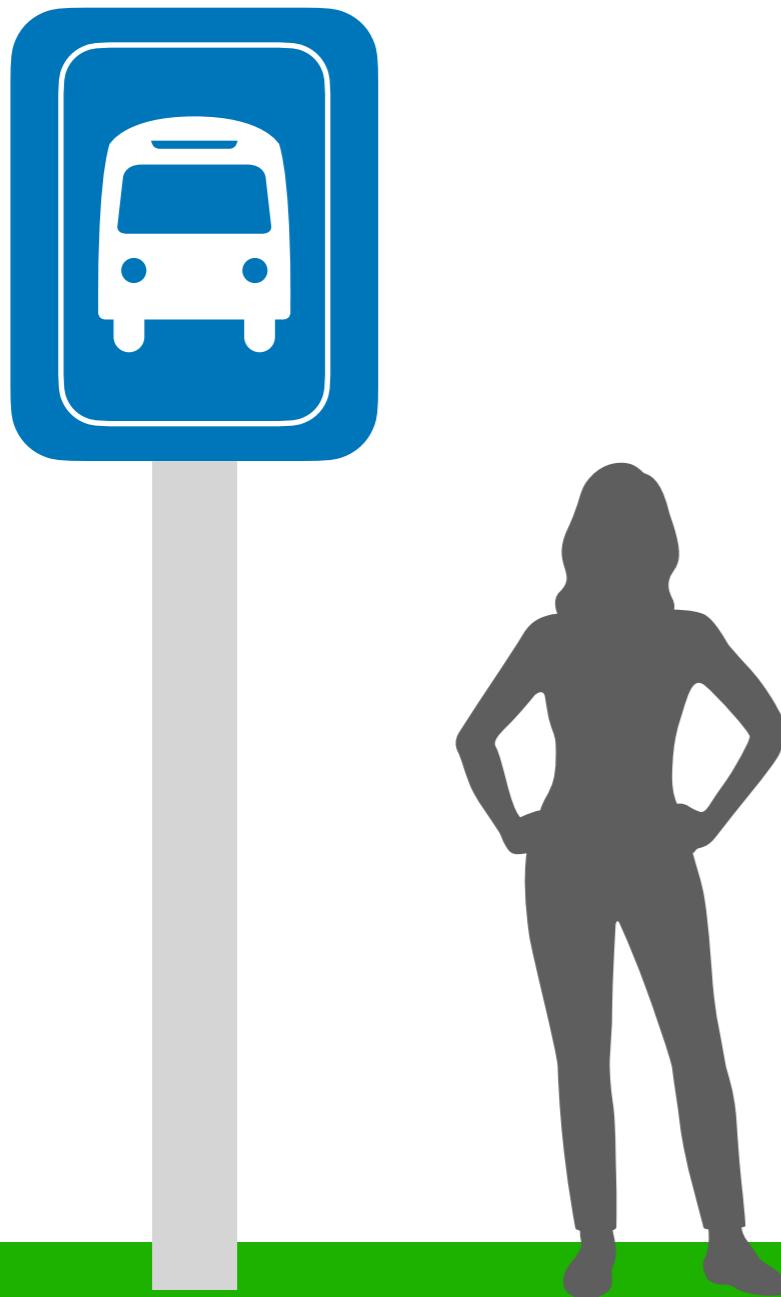
Lecture 9

Models of evolution
Botany 563 – Spring 2022

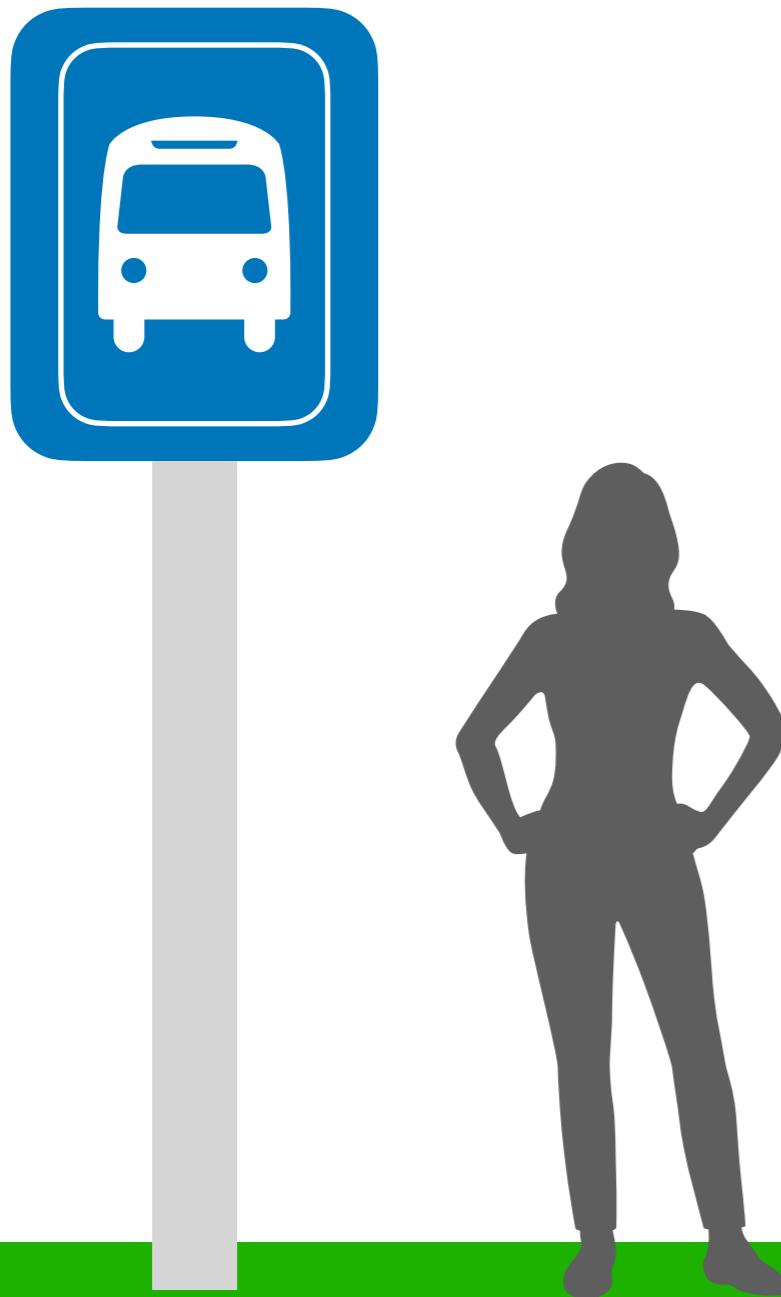
- **Previous class check-up:**
 - We studied the distance and parsimony methods and their strengths and weaknesses
- **Learning Objectives:** At the end of today's session, you will be able to
 - Explain the main characteristics of the substitution model in molecular evolution
 - Assess whether existing literature provide sufficient details on the model assumptions
- **Pre-class work**
 - Read HAL 1.1

What is a model?

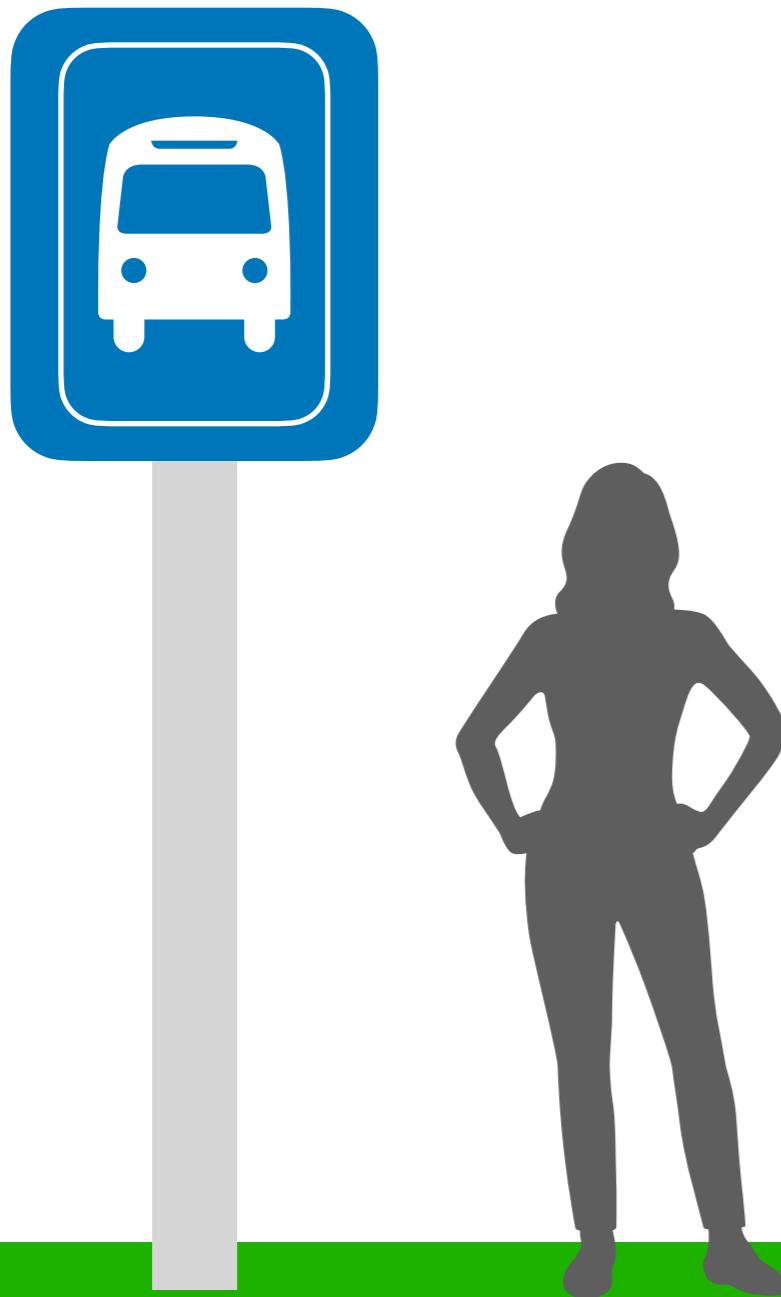
Example



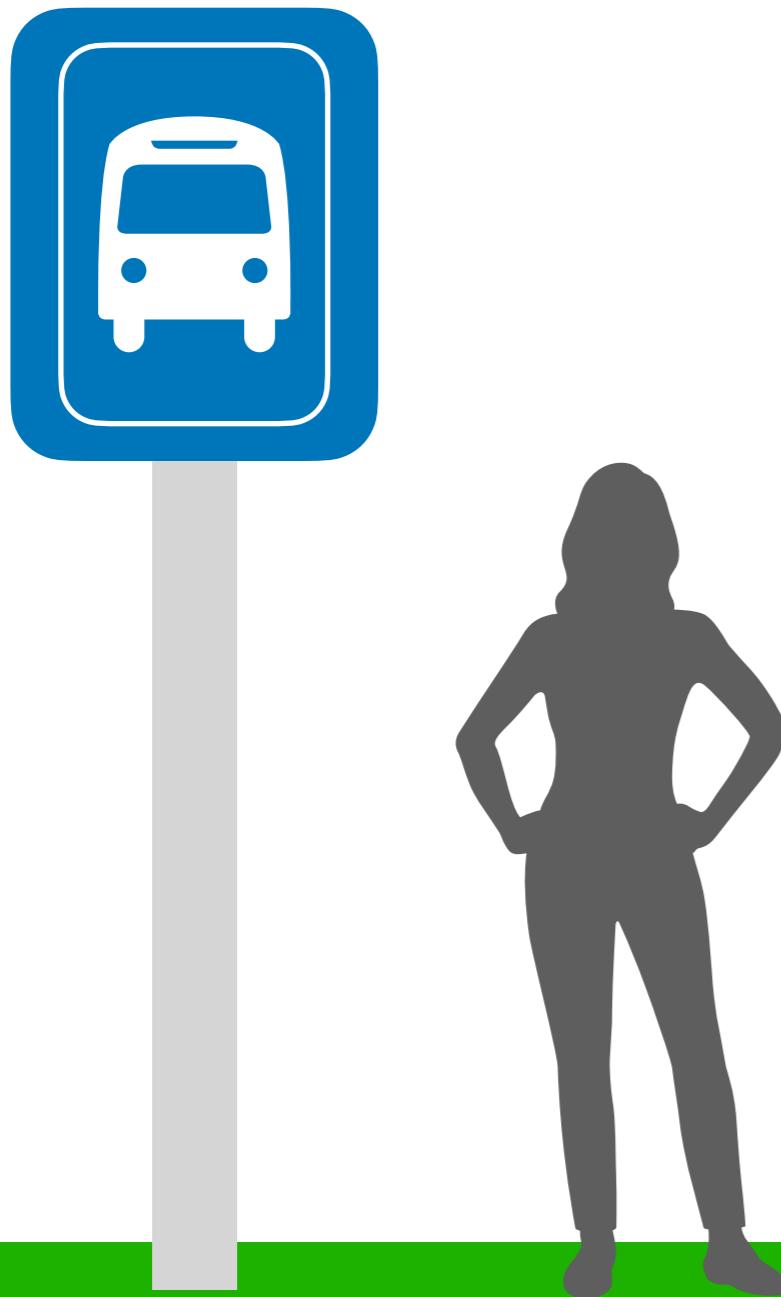
Example



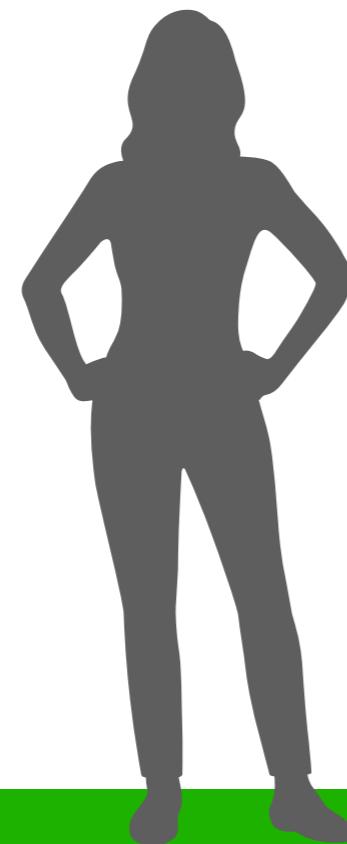
Example



Example

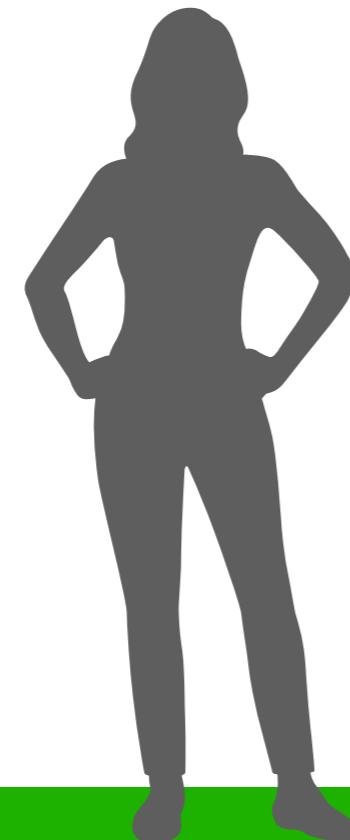


Example



**How can you predict the
number of bikes that you
will see the next day?**

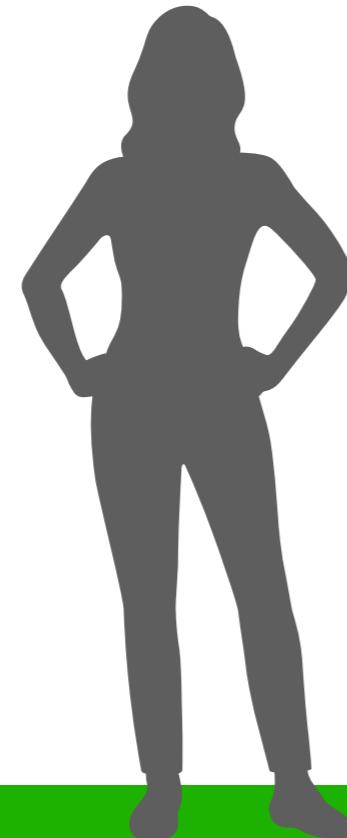
Example



How can you predict the number of bikes that you will see the next day?

Count the number of bikes for n days and then use the average to predict

Example



How can you predict the number of bikes that you will see the next day?

Count the number of bikes for n days and then use the average to predict

The process is random, so you don't expect to be correct, you only get a ballpark

Example



**What if you want to
estimate the probability
that you will see 5 bikes
or 0 bikes?**

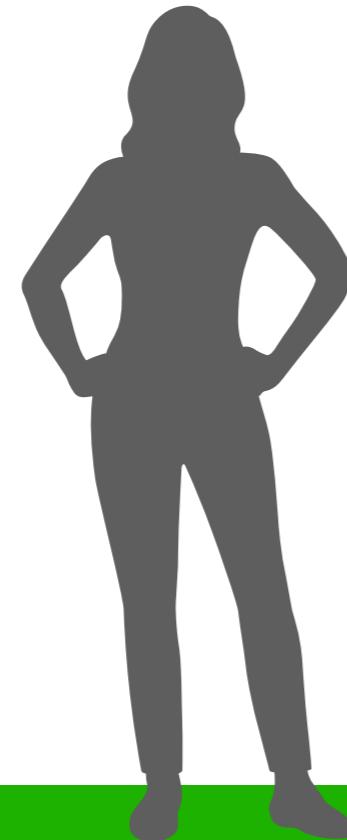
Example



**What if you want to
estimate the probability
that you will see 5 bikes
or 0 bikes?**

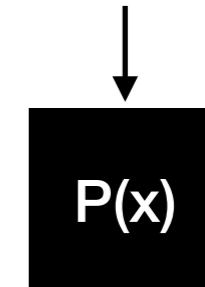
Now you need a probability model!

Example

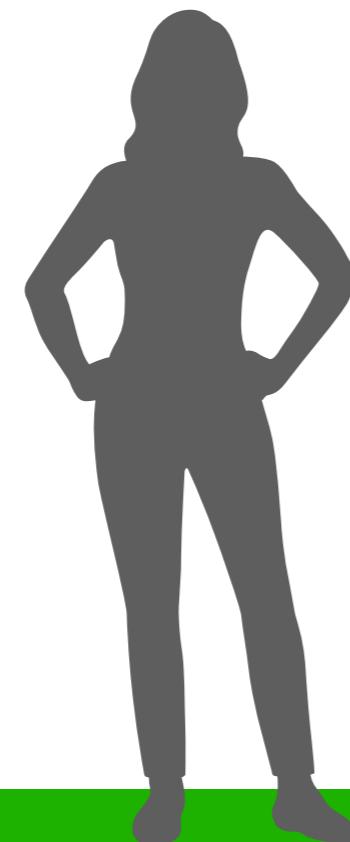


**What if you want to
estimate the probability
that you will see 5 bikes
or 0 bikes?**

Now you need a probability model!

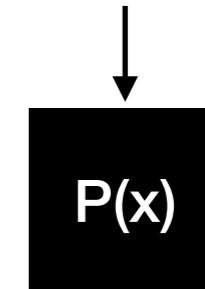


Example



**What if you want to
estimate the probability
that you will see 5 bikes
or 0 bikes?**

Now you need a probability model!



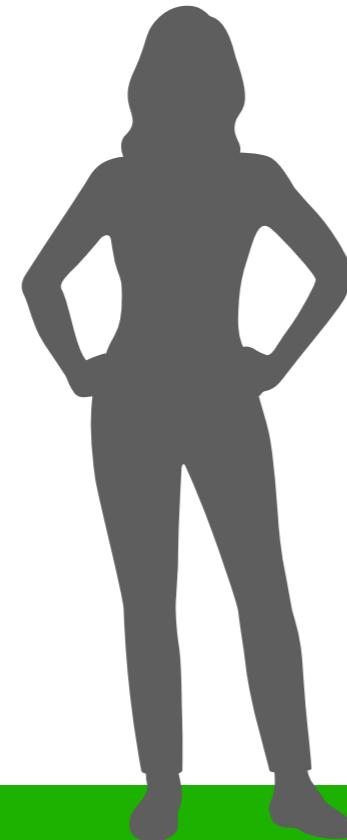
$P(5)$: probability of observing 5 bikes
 $P(0)$: probability of observing 0 bikes

Example



Now you need a probability model!

Example



Now you need a probability model!

Assumptions:

- Symmetric around the mean?
- Not symmetric around the mean?
- How variable?

Example



Now you need a probability model!

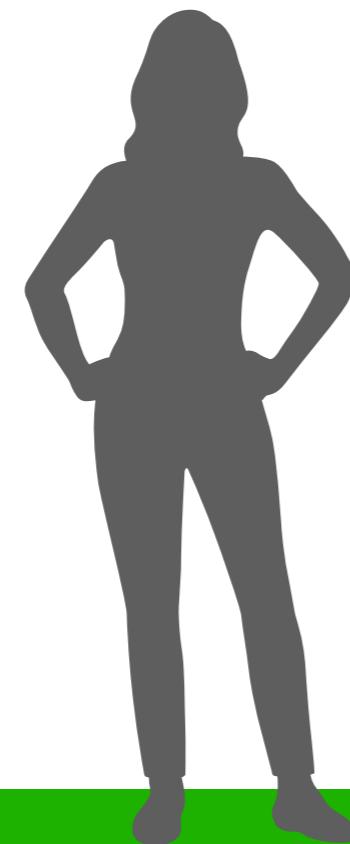
Example



Now you need a probability model!

You investigate that the Poisson model
is a widely used model for counts

Example



Now you need a probability model!

You investigate that the Poisson model
is a widely used model for counts

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Example



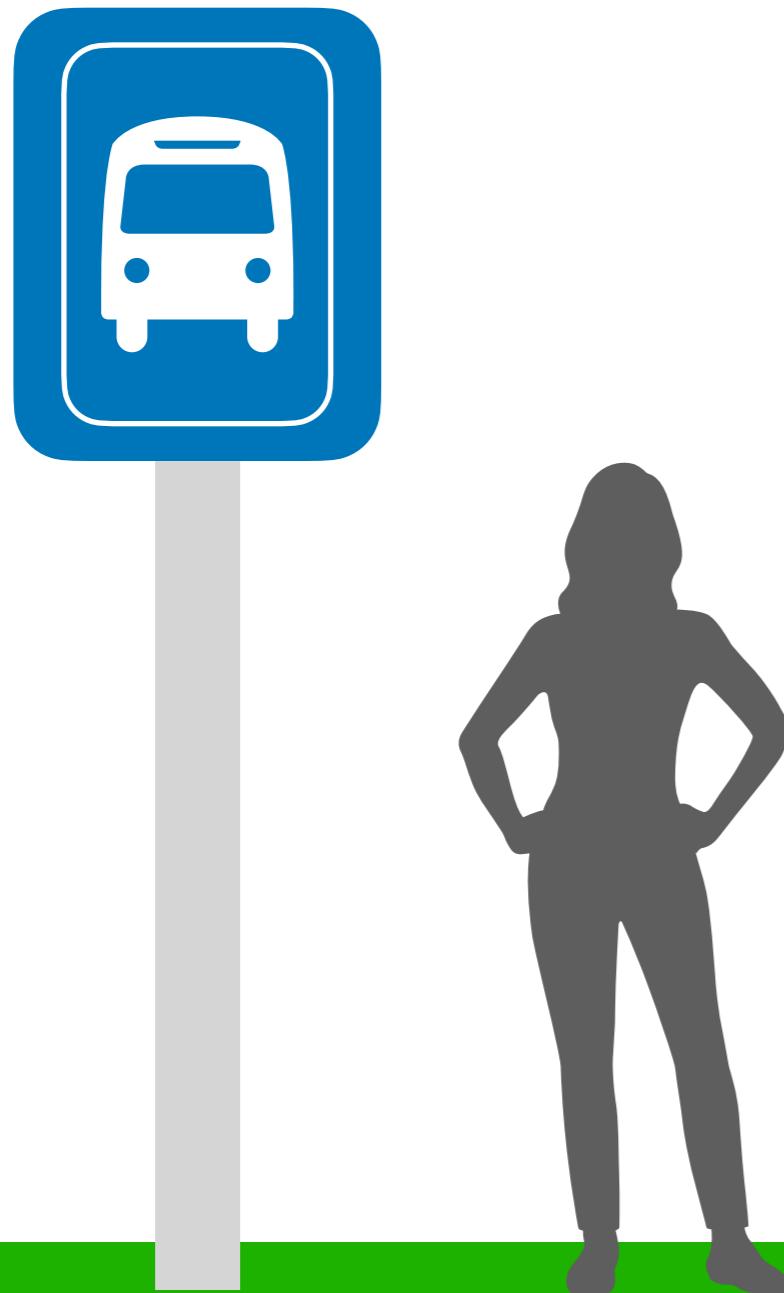
Now you need a probability model!

You investigate that the Poisson model
is a widely used model for counts

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Average

Example



Now you need a probability model!

You investigate that the Poisson model
is a widely used model for counts

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

A yellow box labeled "Average" with an arrow points to the λ term in the equation.

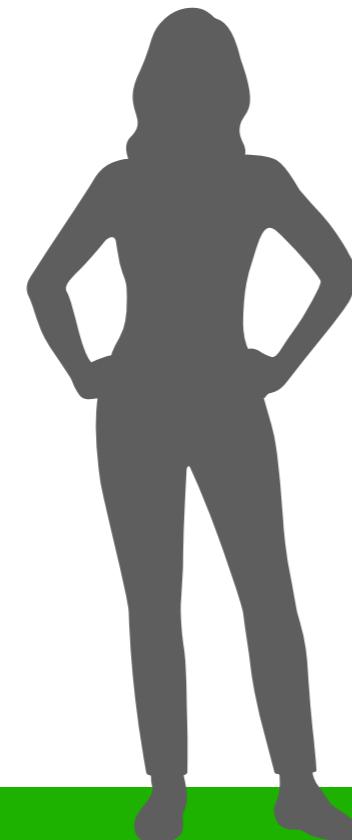
Let's say that the average
number of bikes you see per
day is 8.4

$$P(X=5) = 0.07837$$

$$P(X=0) = 0.00022$$

<https://stattrek.com/online-calculator/poisson.aspx>

Example



Now you need a probability model!

You investigate that the Poisson model
is a widely used model for counts

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

A yellow box labeled "Average" with an arrow pointing to the λ term in the equation.

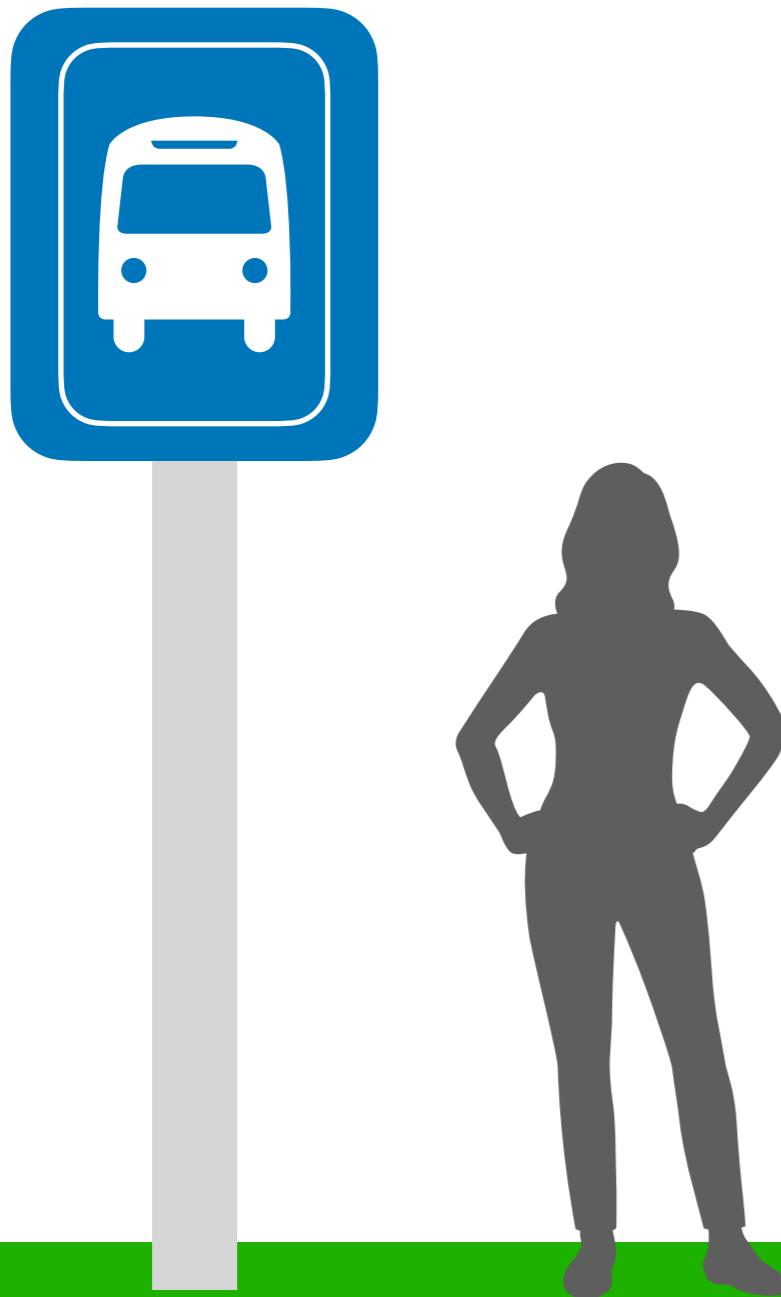
Let's say that the average
number of bikes you see per
day is 8.4

$$P(X > 0) = 0.99978$$

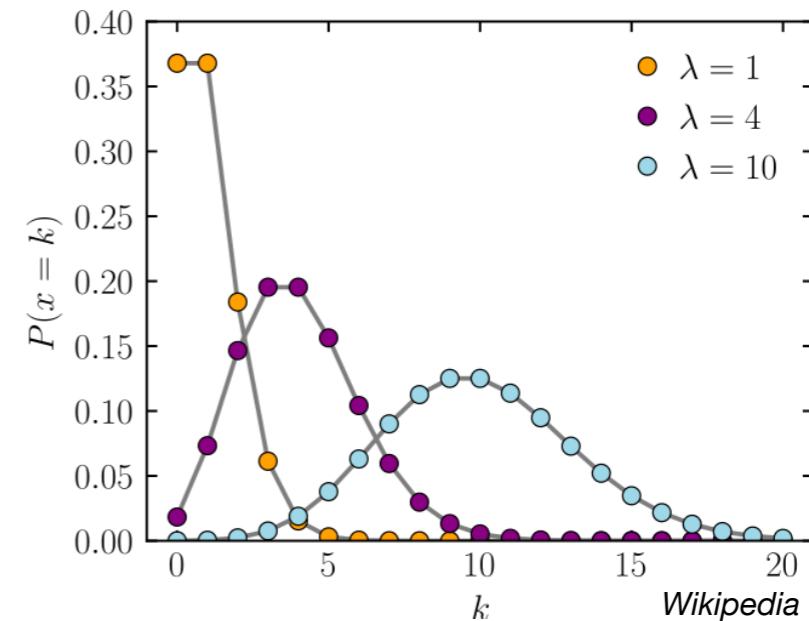
$$P(X > 7) = 0.60135$$

<https://stattrek.com/online-calculator/poisson.aspx>

Example



- We have a probability model (Poisson) for the number of bikes you see every day
- Your model only has one parameter (λ) that governs the mean and shape of the distribution



- You use the observed average number of bikes to estimate the lambda parameter of the model

Example

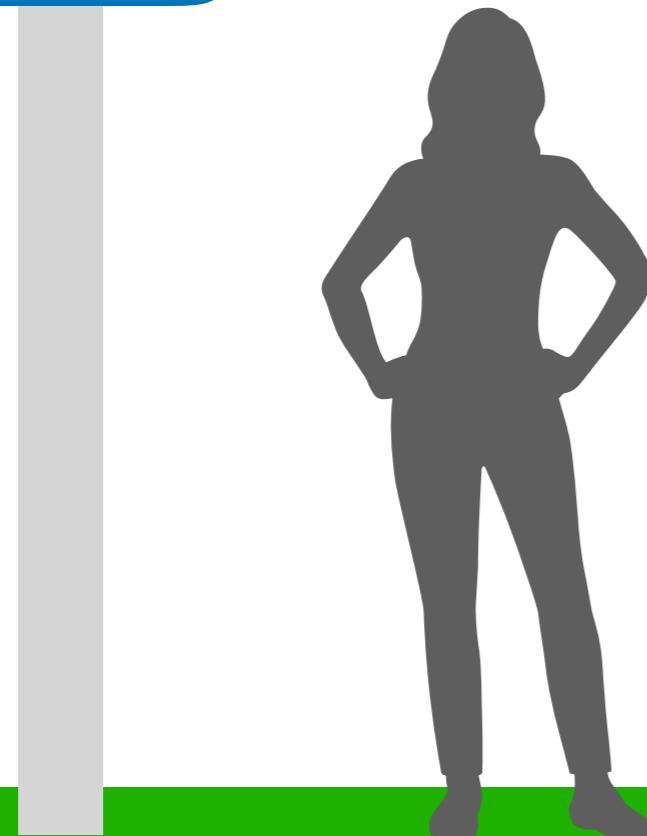
- We have a probability model (Poisson) for the number of bikes you see every day



Questions to ask:

- Does the model capture the mean behavior?
- Does the model capture the variability?
- Can we accurately get probabilities of extreme events: $P(X=1000)$?

Example



- We have a probability model (Poisson) for the number of bikes you see every day

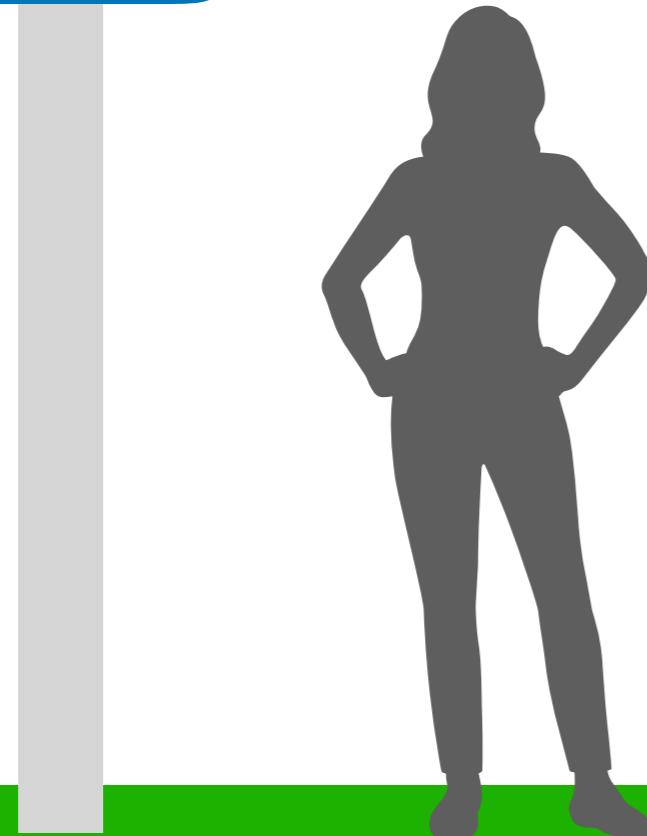
Questions to ask:

- Does the model capture the mean behavior?
- Does the model capture the variability?
- Can we accurately get probabilities of extreme events: $P(X=1000)$?

Even if it is a good model, it will never be perfect:

- Only one parameter controls mean and shape of distribution
- We assume the same parameter for every day and for every hour of the day:
 - Day to day variability: not captured
 - Hour to hour variability: not captured
 - Weather: not captured

Example



- We have a probability model (Poisson) for the number of bikes you see every day

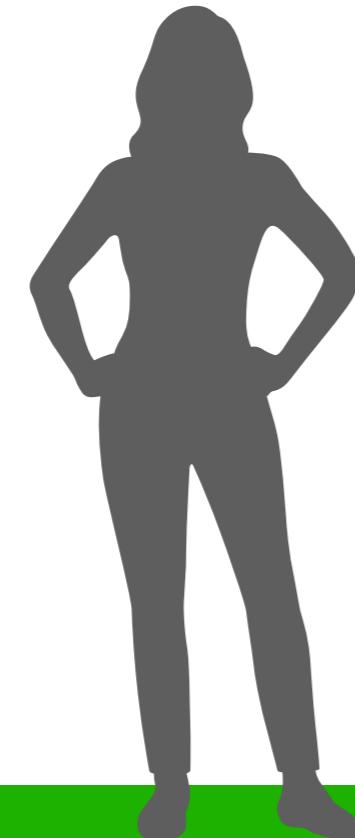
Questions to ask:

- Does the model capture the mean behavior?
- Does the model capture the variability?
- Can we accurately get probabilities of extreme events: $P(X=1000)$?

Even if it is a good model, it will never be perfect:

- Only one parameter controls mean and shape of distribution
- We assume the same parameter for every day and for every hour of the day:
 - Day to day variability: not captured
 - Hour to hour variability: not captured
 - Weather: not captured

Example



- We have a probability model (Poisson) for the number of bikes you see every day

Questions to ask:

- Does the model capture the mean behavior?
- Does the model capture the variability?
- Can we accurately get probabilities of extreme events: $P(X=1000)$?

It does not have to!

Even if it is a good model, it will never be perfect:

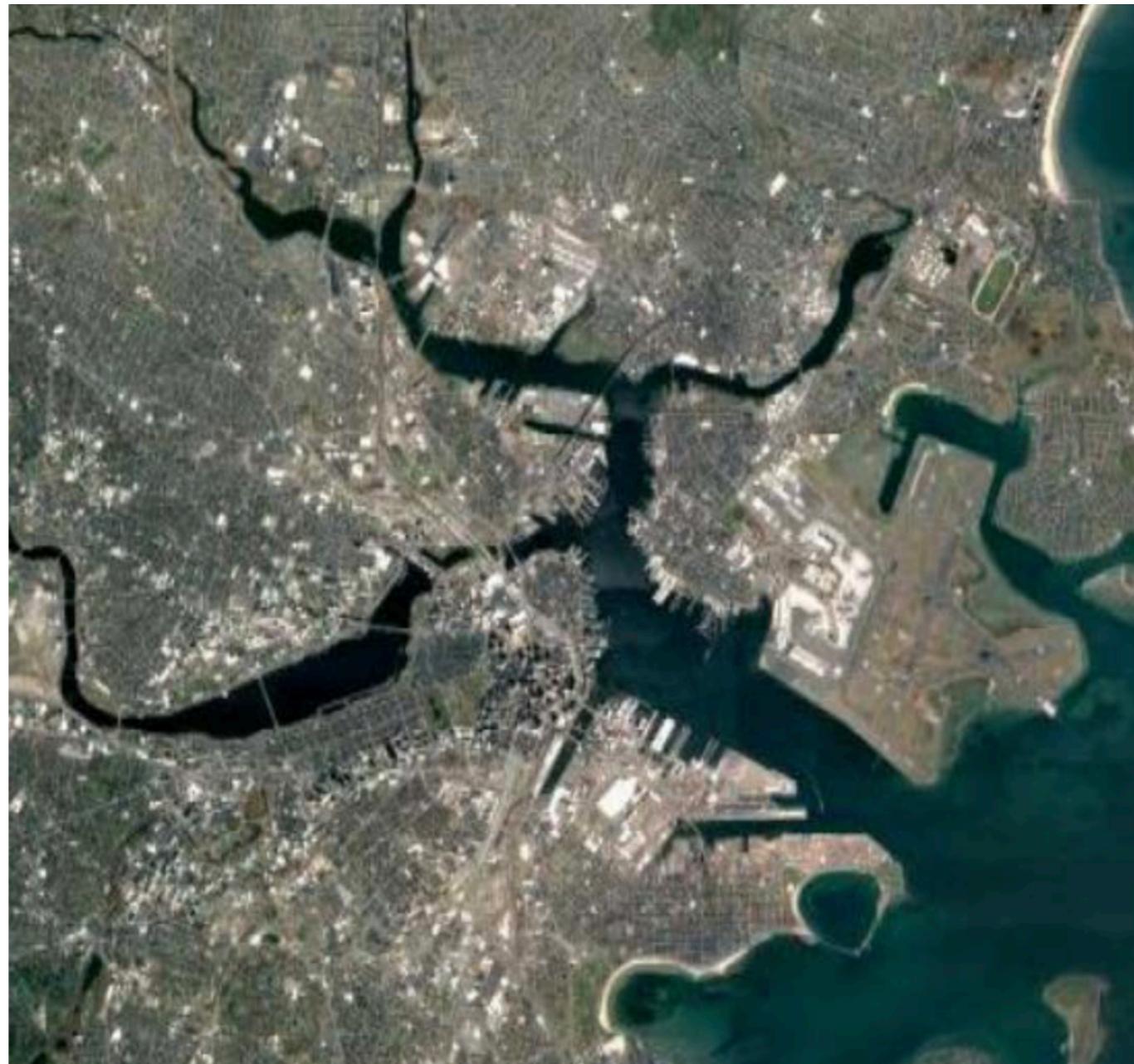
- Only one parameter controls mean and shape of distribution
- We assume the same parameter for every day and for every hour of the day:
 - Day to day variability: not captured
 - Hour to hour variability: not captured
 - Weather: not captured

**All models are wrong,
but some models are
useful**



George Box
Founder of UW Stat
department

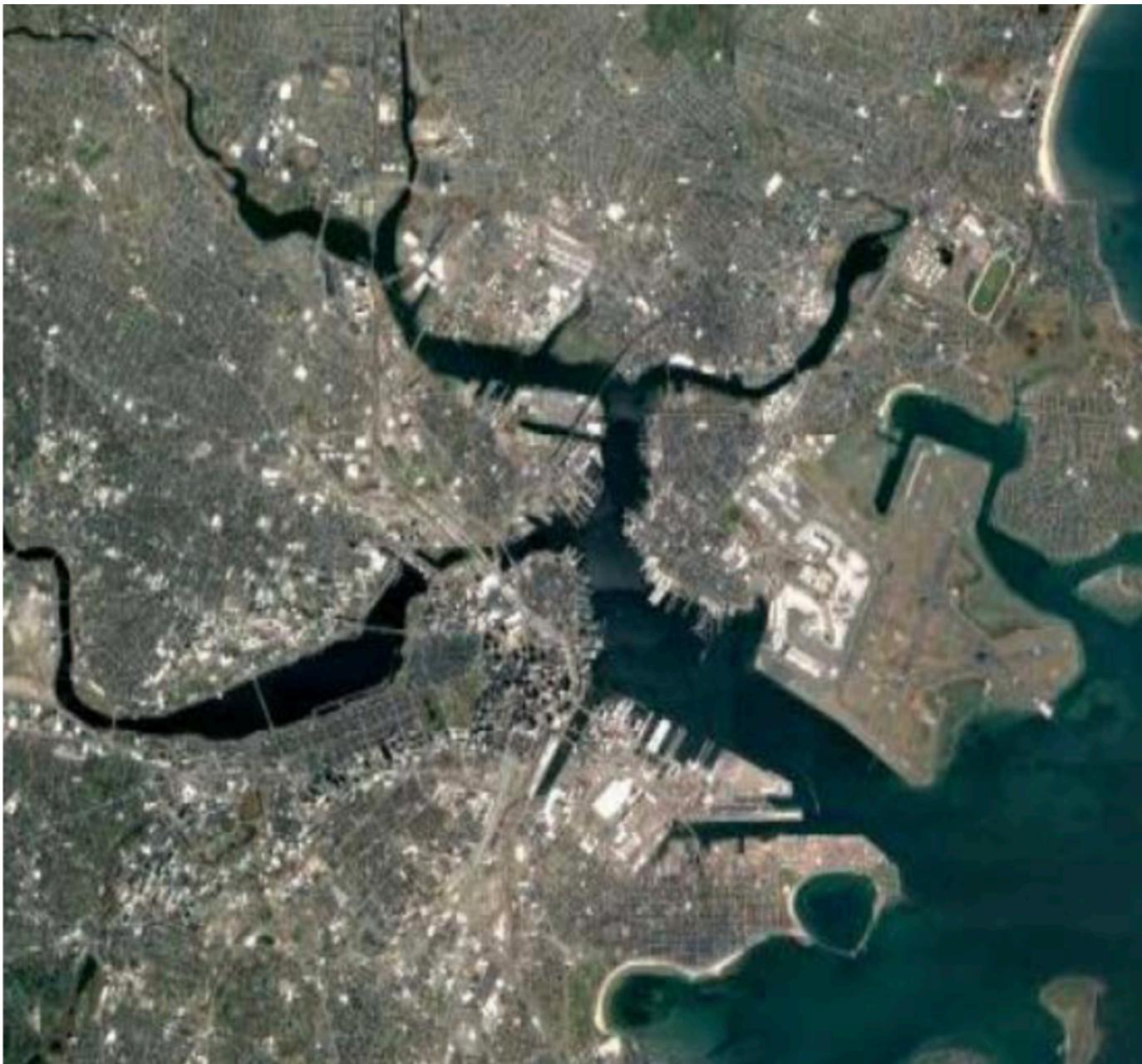
A very **realistic** subway map



A very practical subway map



Which one is more useful?



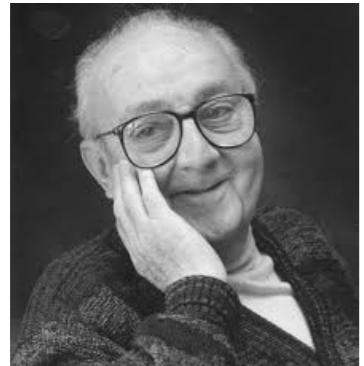
**All models are wrong,
but some models are
useful**



George Box
Founder of UW Stat
department

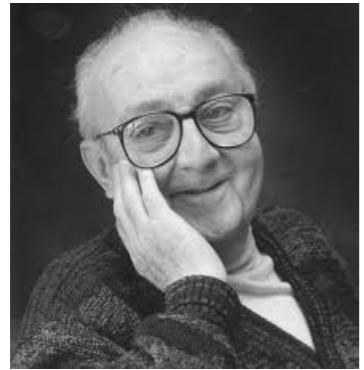
All models are wrong,
but some models are
useful

How useful?



George Box
Founder of UW Stat
department

**All models are wrong,
but some models are
useful**

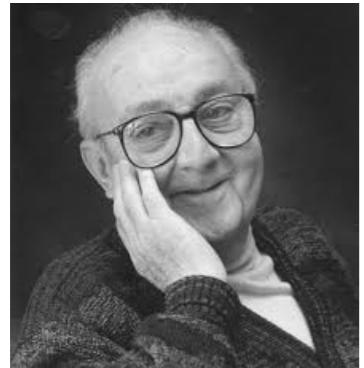


George Box
Founder of UW Stat
department

How useful?

It will depend on the
assumptions of the model
and the quality of the
input data

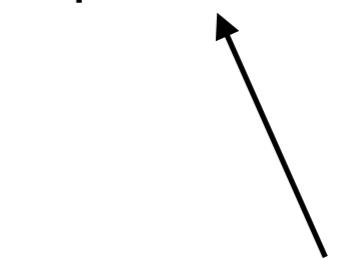
**All models are wrong,
but some models are
useful**



George Box
Founder of UW Stat
department

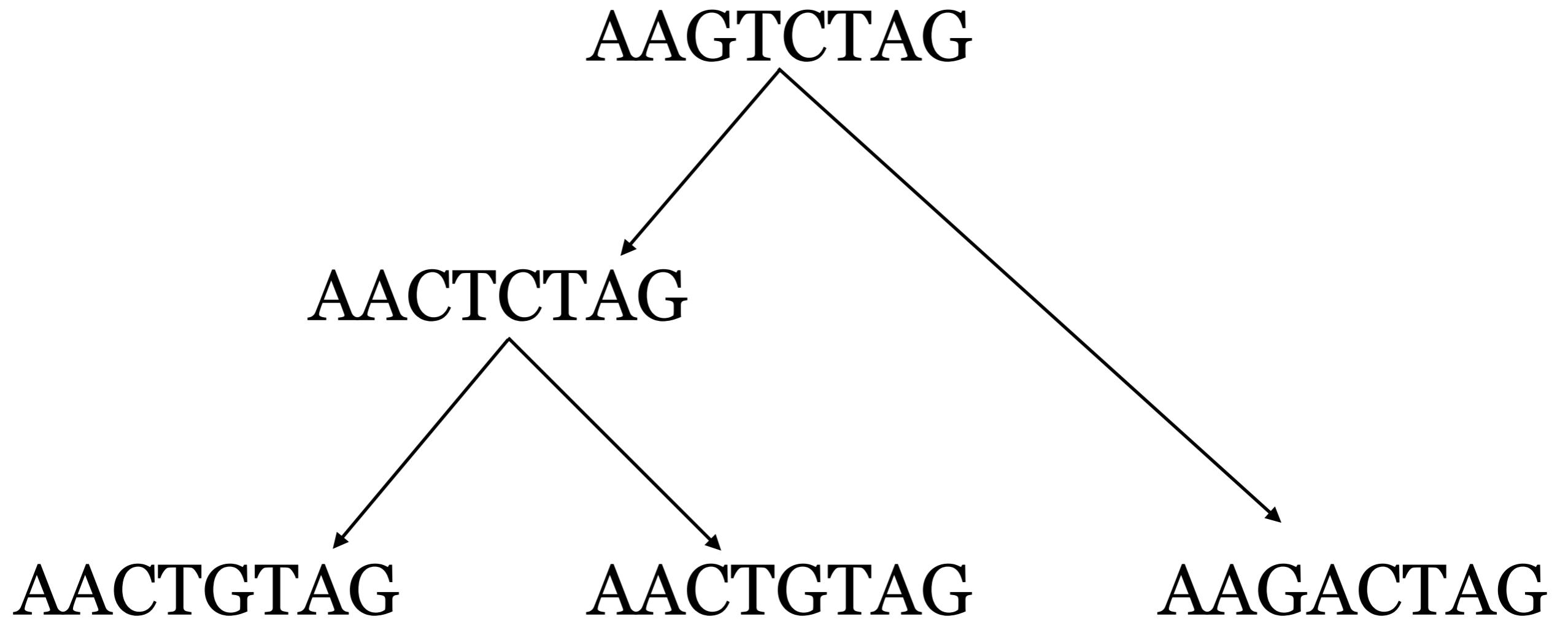
How useful?

It will depend on the
assumptions of the model
and the quality of the
input data

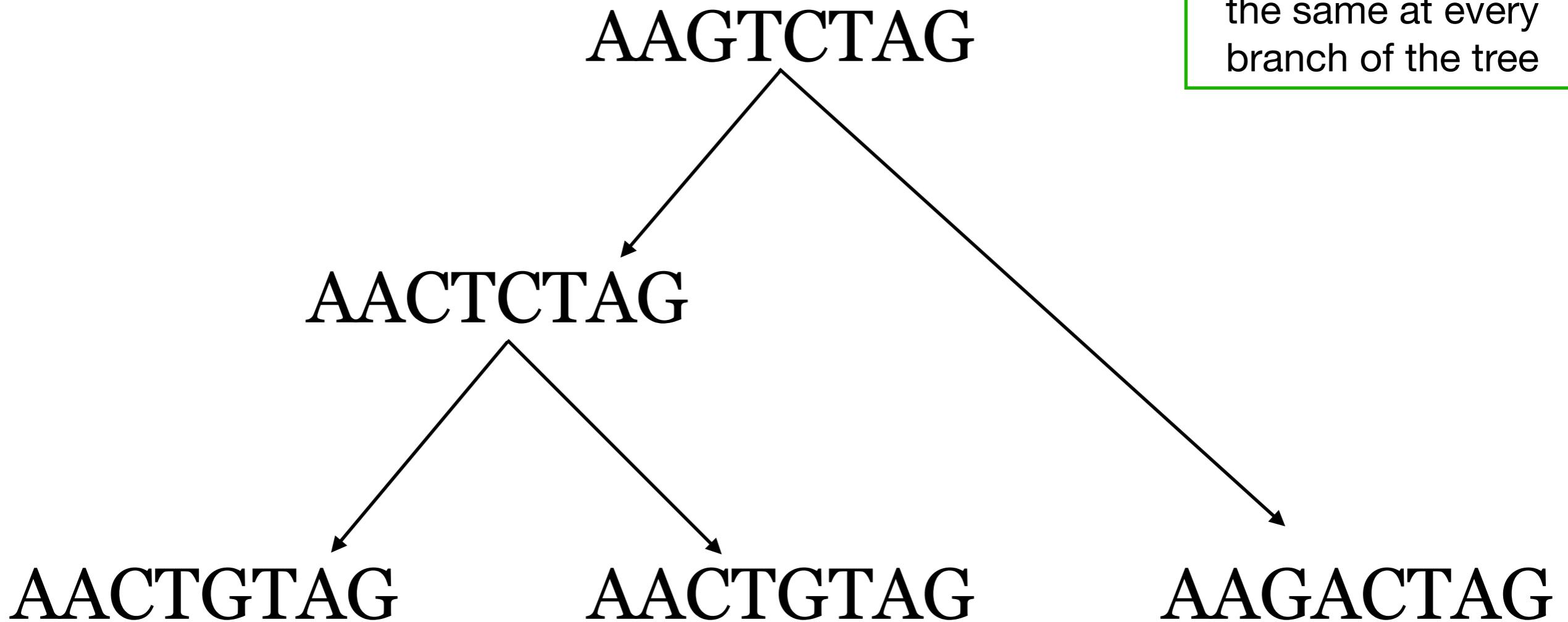


In our example, we count the
number of bikes on n days

We need a probability model for the evolution of sequences along a phylogenetic tree



We need a probability model for the evolution of sequences along a phylogenetic tree



Assumption 1: The mutation process is the same at every branch of the tree

We need a probability model for the evolution of one ancestral sequence into a descendant sequence

AAGTCTAG



AACTCTAG

Assumption 1: The mutation process is the same at every branch of the tree



Implication: We only focus on the mutation process between two sequences

We need a probability model for the evolution of one ancestral sequence into a descendant sequence

Assumption 2: We assume sites evolve independently

AAGTCTAG



AACTCTAG

We need a probability model for the evolution of an ancestral site into a descendant site

Assumption 2: We assume sites evolve independently

Implication: We can focus on the mutation process between two sites

A



A

We need a probability model for the evolution of an ancestral site into a descendant site

A



A

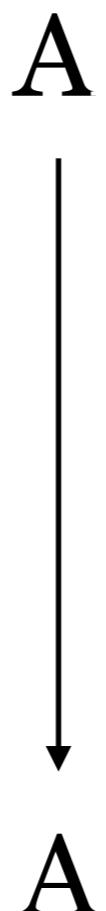
Assumption 2: We assume sites evolve independently



Implication: We can focus on the mutation process between two sites

Assumption 3: All sites evolve the same

We need a probability model for the evolution of an ancestral site into a descendant site



Assumption 2: We assume sites evolve independently

Implication: We can focus on the mutation process between two sites

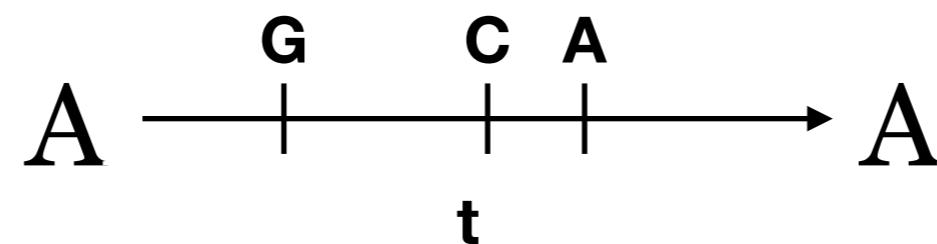
Assumption 3: All sites evolve the same

Implication: We can choose any site to model the mutation process

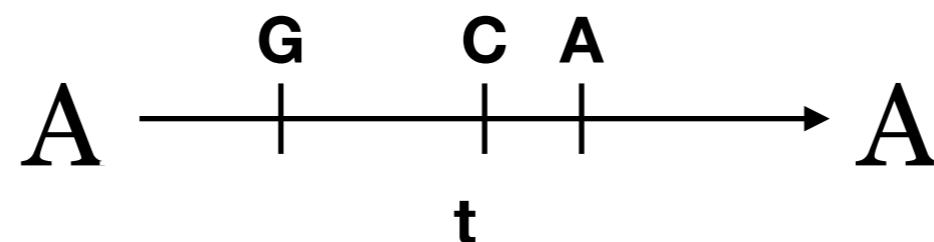
We need a probability model for the evolution of an ancestral site into a descendant site

$$A \xrightarrow{t} A$$

We need a probability model for the evolution of an ancestral site into a descendant site



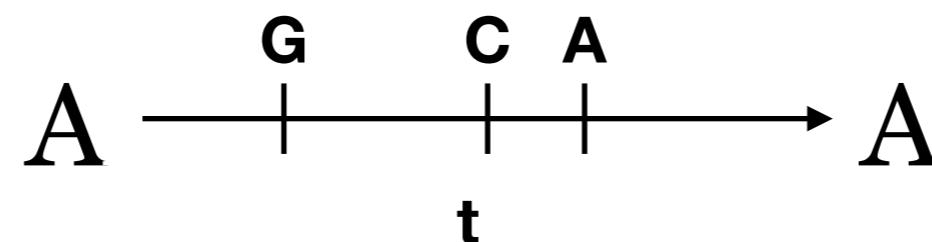
We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

We need a probability model for the evolution of an ancestral site into a descendant site

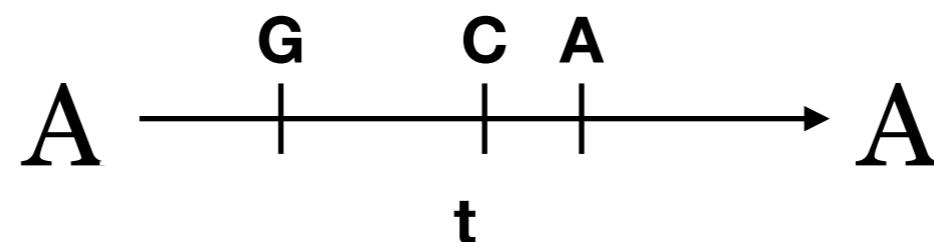


Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

Rate of mutation events

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

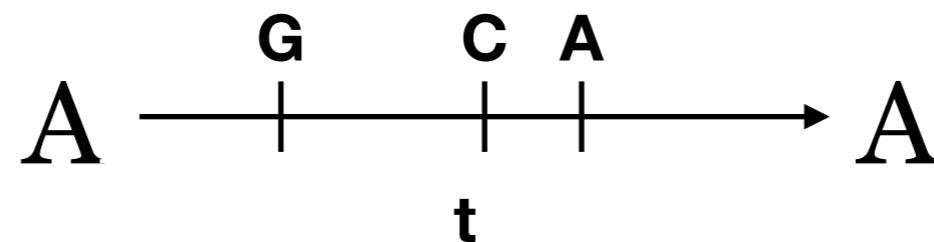
$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

Rate of mutation events

Expected number of mutation events in time t

The equation is $P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$. A green circle highlights $(\mu t)^k$, which is connected by a dotted green line to the text 'Expected number of mutation events in time t'. A red circle highlights $e^{-\mu t}$, which is connected by a dotted red line to the text 'Rate of mutation events'.

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

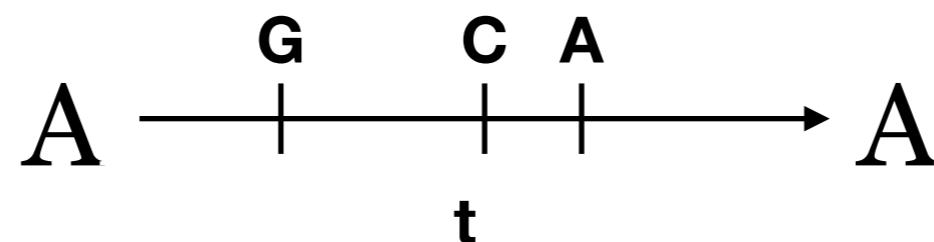
Rate of mutation events

Expected number of mutation events in time t

The equation is a Poisson distribution formula. The term $(\mu t)^k$ is highlighted with a green circle and connected by a dotted green line to the text "Expected number of mutation events in time t ". The term $e^{-\mu t}$ is highlighted with a red circle and connected by a dotted red line to the text "Rate of mutation events".

* Note that “mutation events” are not the same as substitutions because A->A is considered a mutation event

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

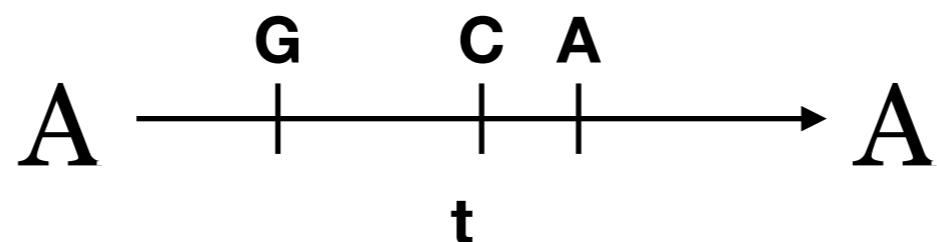
$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

Probability matrix $R =$

	A	C	G	T
A				
C				
G				
T				

Probability of changing from state A to state T

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

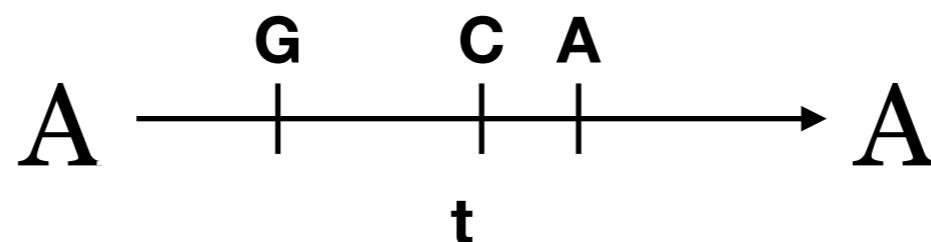
$$\mathbf{R} =$$

	A	C	G	T
A				
C				
G				
T				

Probability matrix $\mathbf{P}(t) =$

	A	C	G	T
A				
C				
G				
T				

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

$$\mathbf{R} =$$

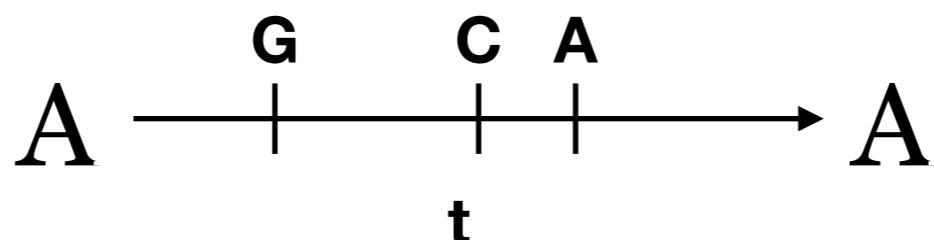
	A	C	G	T
A				
C				
G				
T				

Probability matrix $\mathbf{P}(t) =$

	A	C	G	T
A				
C				
G				
T				

Probability of starting on A and ending on T in time t

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

$$\mathbf{R} =$$

	A	C	G	T
A				
C				
G				
T				

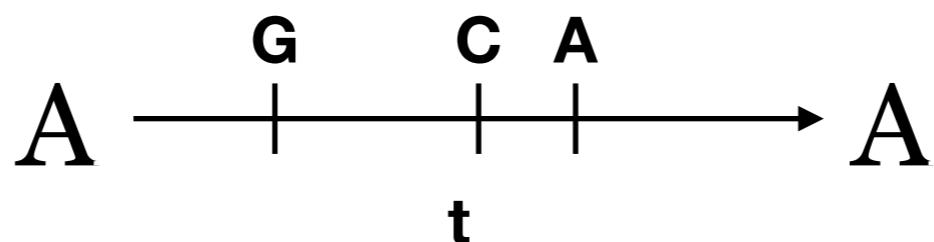
Probability matrix $\mathbf{P}(t) =$

	A	C	G	T
A				
C				
G				
T				

$$= \sum_{k=0}^{\infty} (\mathbf{R}^k) \frac{(\mu t)^k e^{-\mu t}}{k!}$$

Probabilities of change summed over all possible values of k (number of events)

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

$$\mathbf{R} =$$

	A	C	G	T
A				
C				
G				
T				

After some matrix algebra

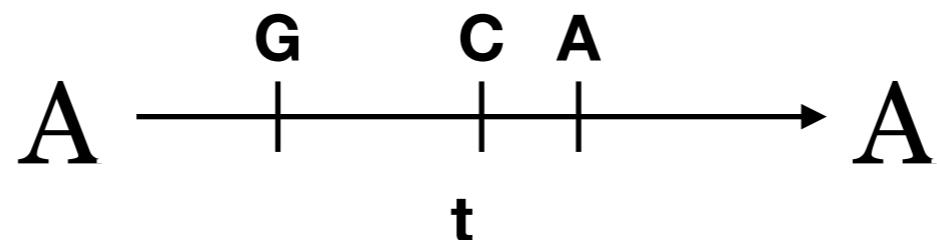
$$e^{\mathbf{Q}\mu t} \stackrel{\downarrow}{=} \mathbf{P}(t) =$$

	A	C	G	T
A				
C				
G				
T				

$$= \sum_{k=0}^{\infty} (\mathbf{R}^k) \frac{(\mu t)^k e^{-\mu t}}{k!}$$

Probabilities of change summed over all possible values of k (number of events)

We need a probability model for the evolution of an ancestral site into a descendant site



Number of mutations on time t is assumed to follow a Poisson distribution

$$P(X = k) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

$$\mathbf{R} =$$

	A	C	G	T
A				
C				
G				
T				

After some matrix algebra

$$e^{\mathbf{Q}ut} \stackrel{\downarrow}{=} \mathbf{P}(t) =$$

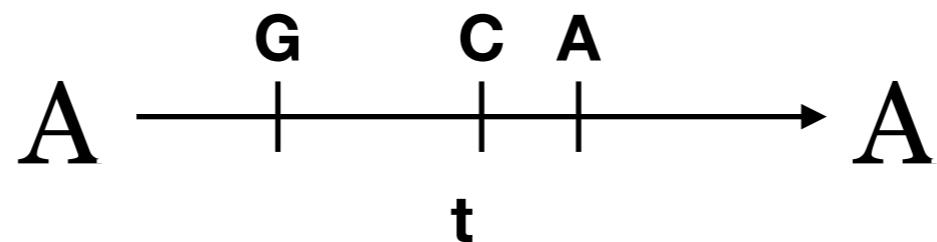
	A	C	G	T
A				
C				
G				
T				

Instantaneous rate matrix
(generator)
 $\mathbf{Q} = \mathbf{R} - \mathbf{I}$

$$= \sum_{k=0}^{\infty} (\mathbf{R}^k) \frac{(\mu t)^k e^{-\mu t}}{k!}$$

Probabilities of change summed over all possible values of k (number of events)

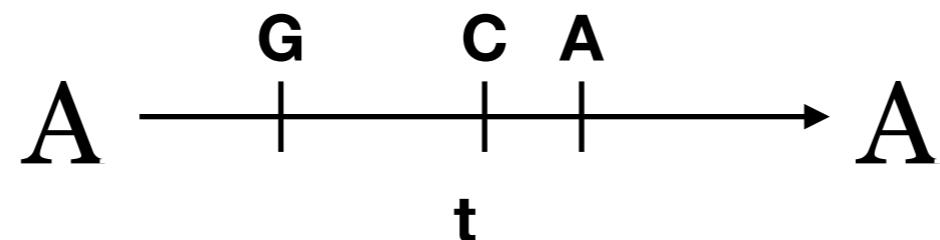
We need a probability model for the evolution of an ancestral site into a descendant site



$$\mathbf{P}(t) = \begin{matrix} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & & & & \\ \text{C} & & & & \\ \text{G} & & & & \\ \text{T} & & & & \end{matrix} = e^{\mathbf{Q}\mu t}$$

Substitution model

We need a probability model for the evolution of an ancestral site into a descendant site



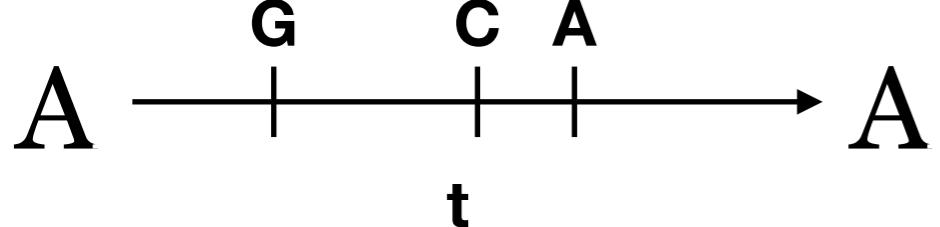
$$P(t) = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{matrix} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{matrix} \end{matrix} = e^{Q\mu t}$$

Substitution model

Choosing a substitution model means choosing Q:

$$Q = \begin{pmatrix} -q_A & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & -q_C & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & -q_G & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & -q_T \end{pmatrix}$$

We need a probability model for the evolution of an ancestral site into a descendant site

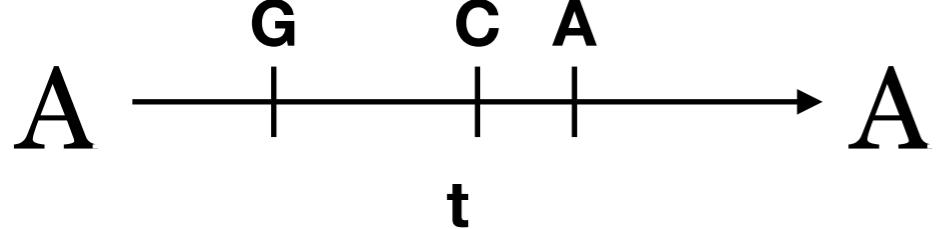


$$P(t) = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{matrix} & & & \\ & & & \\ & & & \\ & & & \end{matrix} \end{matrix} = e^{\mathbf{Q}\mu t}$$
$$\mathbf{Q} = \begin{pmatrix} -q_A & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & -q_C & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & -q_G & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & -q_T \end{pmatrix}$$

Substitution model

**Homogeneous Continuous-time Markov chain
on 4 states: A,C,G,T**

We need a probability model for the evolution of an ancestral site into a descendant site



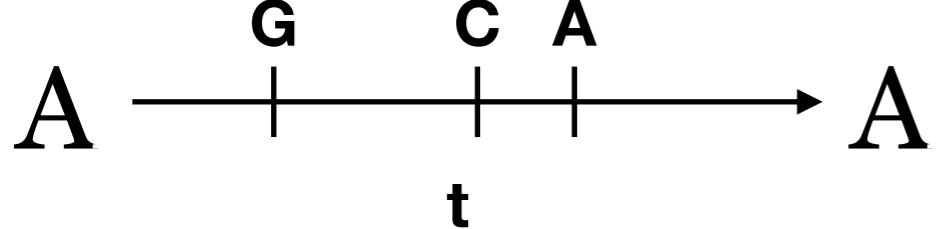
$$P(t) = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{matrix} & & & \\ & & & \\ & & & \\ & & & \end{matrix} \end{matrix} = e^{\mathbf{Q}\mu t}$$
$$\mathbf{Q} = \begin{pmatrix} -q_A & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & -q_C & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & -q_G & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & -q_T \end{pmatrix}$$

Substitution model

Homogeneous Continuous-time Markov chain

Probabilities for the next event do not depend on the time point where the chain is

We need a probability model for the evolution of an ancestral site into a descendant site



$$P(t) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \end{array} = e^{\mathbf{Q}\mu t}$$
$$\mathbf{Q} = \begin{pmatrix} -q_A & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & -q_C & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & -q_G & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & -q_T \end{pmatrix}$$

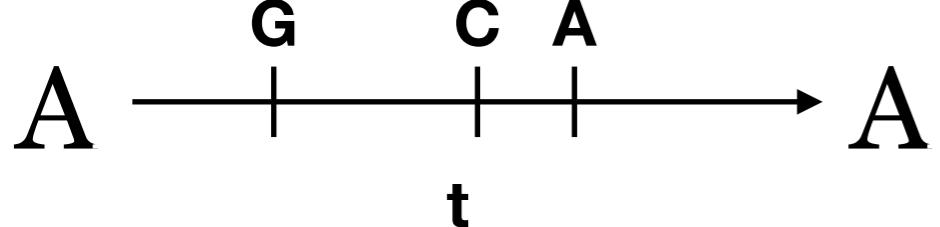
Substitution model

Homogeneous Continuous-time Markov chain

Probabilities for the next event do not depend on the time point where the chain is

Event can happen at any time point (not just discrete steps)

We need a probability model for the evolution of an ancestral site into a descendant site



$$P(t) = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{matrix} & & & \\ & & & \\ & & & \\ & & & \end{matrix} \end{matrix} = e^{\mathbf{Q}\mu t}$$
$$\mathbf{Q} = \begin{pmatrix} -q_A & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & -q_C & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & -q_G & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & -q_T \end{pmatrix}$$

Substitution model

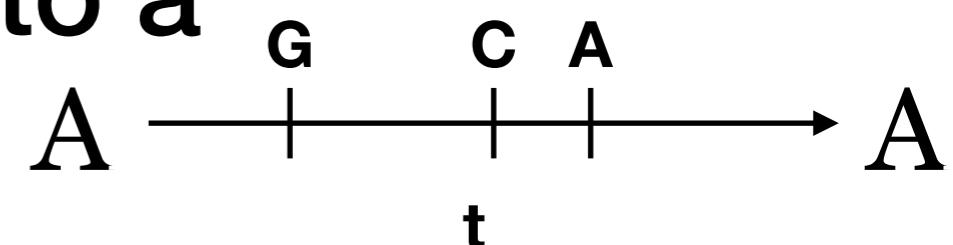
Homogeneous Continuous-time **Markov chain**

Probabilities for the next event do not depend on the time point where the chain is

Event can happen at any time point (not just discrete steps)

Probabilities of the next state only depend on the current state

We need a probability model for the evolution of an ancestral site into a descendant site



Homogeneous Continuous-time Markov chain

$$P(t) = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \end{matrix} = e^{Qt\mu t}$$

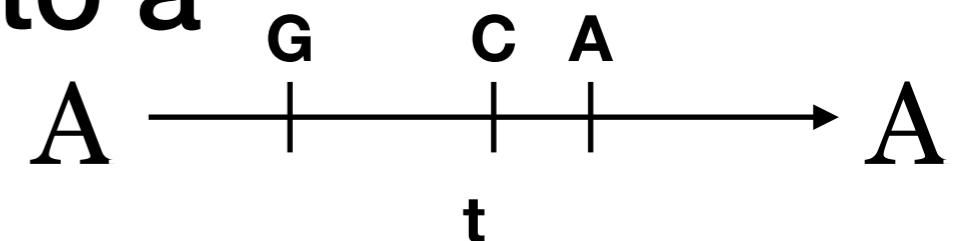
Substitution model

Probabilities for the next event do not depend on the time point where the chain is

Event can happen at any time point (not just discrete steps)

Probabilities of the next state only depend on the current state

We need a probability model for the evolution of an ancestral site into a descendant site



Homogeneous Continuous-time Markov chain

$$P(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{Qt}$$

Substitution model

Probabilities for the next event do not depend on the time point where the chain is

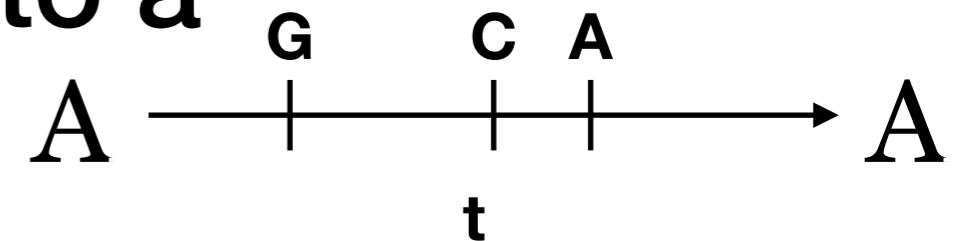
Event can happen at any time point (not just discrete steps)

Probabilities of the next state only depend on the current state

Other mathematical assumptions:

- **Ergodic:** As time goes to infinity, the probability that the site is in some state y is non-zero and independent of the starting state (there is a stationary distribution)
- **Time reversible:** The probability of sampling x from the stationary distribution and going to state y is the same as the probability of sampling y from the stationary distribution and going to state x

We need a probability model for the evolution of an ancestral site into a descendant site



Homogeneous Continuous-time Markov chain

$$P(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{Qt}$$

Substitution model

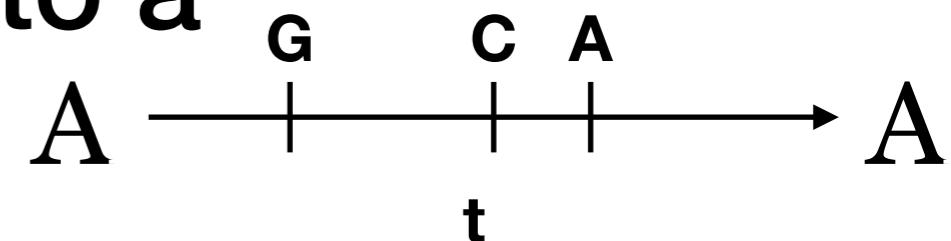
Probabilities for the next event do not depend on the time point where the chain is

Event can happen at any time point (not just discrete steps)

Probabilities of the next state only depend on the current state

Ergodic and time reversible

We need a probability model for the evolution of an ancestral site into a descendant site



Homogeneous Continuous-time Markov chain

$$P(t) = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \end{matrix} = e^{Qt\mu}$$

Substitution model

Probabilities for the next event do not depend on the time point where the chain is

Event can happen at any time point (not just discrete steps)

Probabilities of the next state only depend on the current state

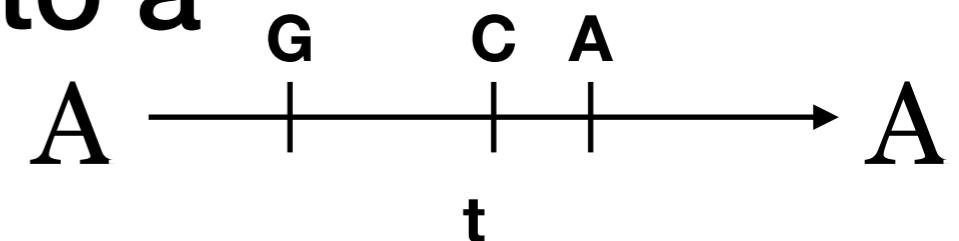
Ergodic and time reversible

Every substitution model that we will study has the same assumptions

Choosing a substitution model means choosing Q:

$$Q = \begin{pmatrix} -q_A & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & -q_C & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & -q_G & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & -q_T \end{pmatrix}$$

We need a probability model for the evolution of an ancestral site into a descendant site



Jukes-Cantor model (JC69)

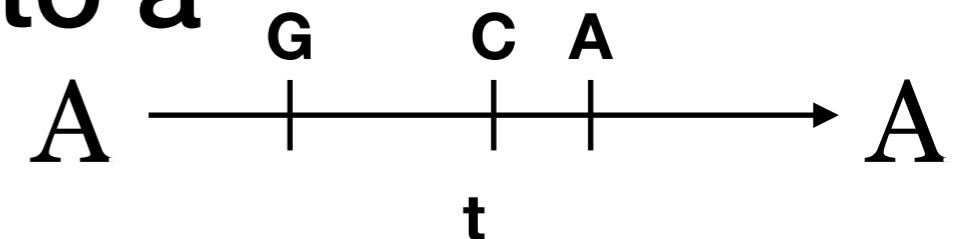
$$\mathbf{Q}_{JC69} = \begin{pmatrix} -3/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & -3/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -3/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & -3/4 \end{pmatrix}$$

$$\pi_A = \pi_C = \pi_G = \pi_T = 1/4$$

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t} \rightarrow P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t\mu}$$

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t\mu}$$

We need a probability model for the evolution of an ancestral site into a descendant site



Felsenstein model (F81)

$$Q_{F81} = \begin{pmatrix} -(\pi_C + \pi_G + \pi_T) & \pi_C & \pi_G & \pi_T \\ \pi_A & -(\pi_A + \pi_G + \pi_T) & \pi_G & \pi_T \\ \pi_A & \pi_C & -(\pi_A + \pi_C + \pi_T) & \pi_T \\ \pi_A & \pi_C & \pi_G & -(\pi_A + \pi_C + \pi_G) \end{pmatrix}$$

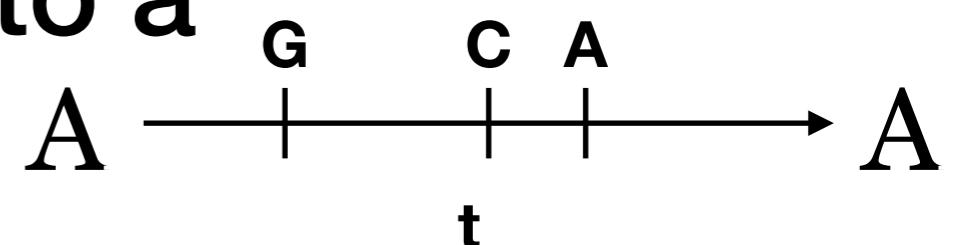
$$\pi_A, \pi_C, \pi_G, \pi_T$$

$$P(t) = e^{Qt} \rightarrow$$

$$P_{ij}(t) = \pi_j(1 - e^{-t\mu})$$

$$P_{ii}(t) = \pi_i + (1 - \pi_i)e^{-t\mu}$$

We need a probability model for the evolution of an ancestral site into a descendant site



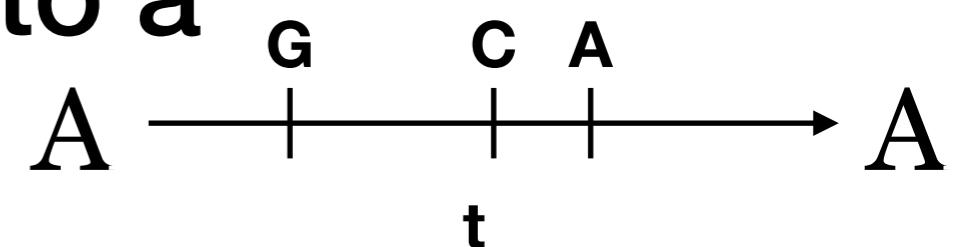
General Time Reversible (GTR)

$$\mathbf{Q}_{GTR} = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

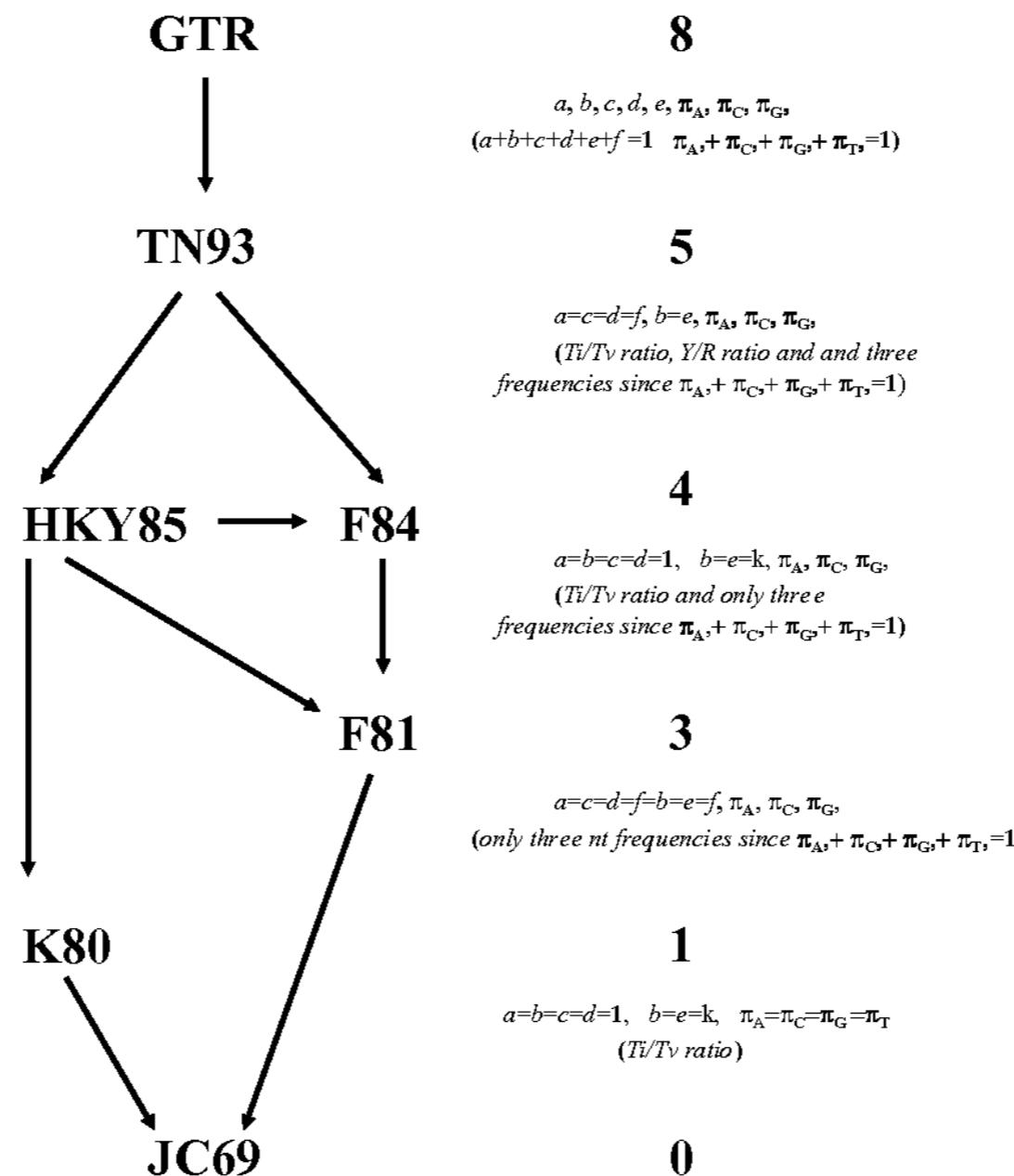
$$\pi_A, \pi_C, \pi_G, \pi_T$$

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

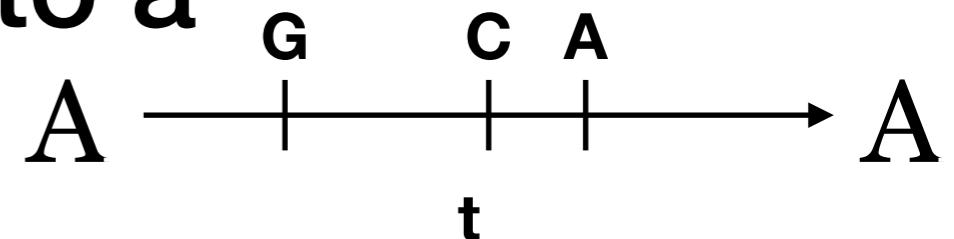
We need a probability model for the evolution of an ancestral site into a descendant site



Model *Free parameters
in the Q-matrix*



We need a probability model for the evolution of an ancestral site into a descendant site



Jukes-Cantor model (JC69)

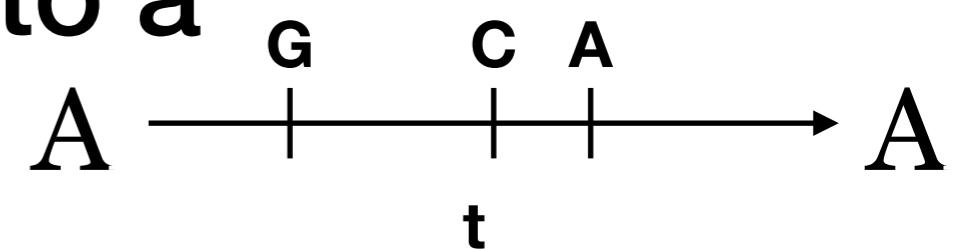
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t\mu}$$

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t\mu}$$

Rate of mutation: Scaling of branch lengths

We need a probability model for the evolution of an ancestral site into a descendant site

Jukes-Cantor model (JC69)

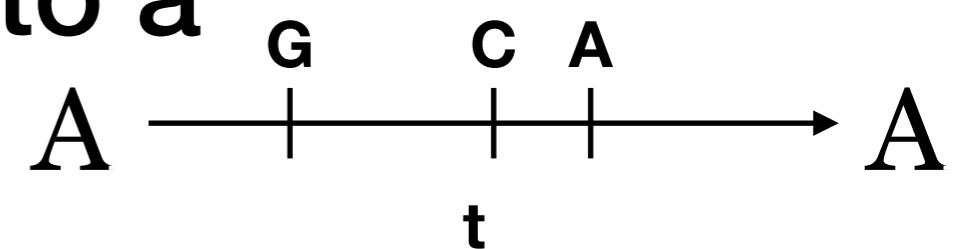


$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t\mu}$$
$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t\mu}$$

Expected number of mutation events in time t

Rate of mutation: Scaling of branch lengths

We need a probability model for the evolution of an ancestral site into a descendant site



Jukes-Cantor model (JC69)

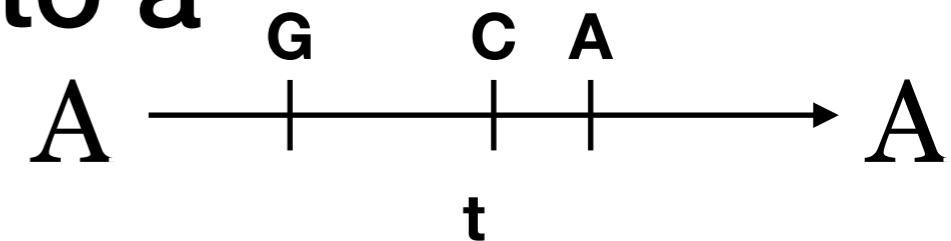
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t\mu}$$
$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t\mu}$$

Expected number of mutation events in time t

Rate of mutation: Scaling of branch lengths

Problem: These mutation events include "redundant events" (A->A)

We need a probability model for the evolution of an ancestral site into a descendant site



Jukes-Cantor model (JC69)

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t\mu}$$
$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t\mu}$$

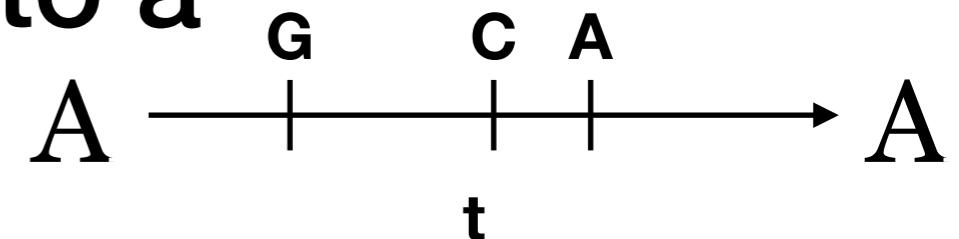
Expected number of mutation events in time t

Rate of mutation: Scaling of branch lengths

Problem: These mutation events include "redundant events" ($A \rightarrow A$)

If we want branch lengths to reflect non-redundant substitutions per site, we need to **scale them appropriately**

We need a probability model for the evolution of an ancestral site into a descendant site



Jukes-Cantor model (JC69)

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t\mu}$$

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t\mu}$$

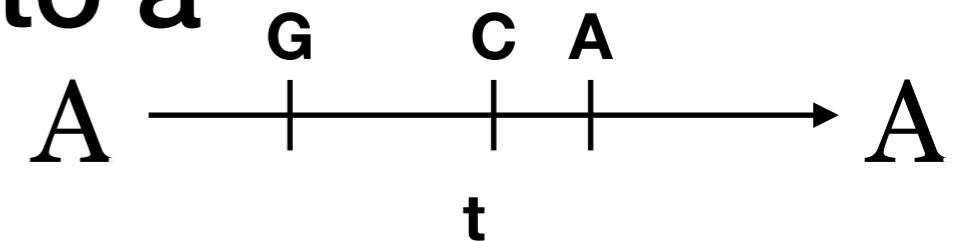
Rate of mutation: Scaling of branch lengths

Instead of having two parameters: \mathbf{Q} , μ ,
for a given \mathbf{Q} , we choose μ such that the overall
rate of mutation is one. In this way, the length of
the branch corresponds to the expected number
of mutations per site along that branch
(irrespective of the model)

More math-y: overall rate of mutation (expected
number of non-redundant events in unit time $t=1$):

$$-\mu \cdot \text{trace}(\Pi \mathbf{Q})$$

We need a probability model for the evolution of an ancestral site into a descendant site



Jukes-Cantor model (JC69)

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t\mu}$$

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t\mu}$$

Rate of mutation: Scaling of branch lengths

Instead of having two parameters: \mathbf{Q}, μ ,

$$-\mu \cdot \text{trace}(\Pi \mathbf{Q}) = 1$$

for a given \mathbf{Q} , we choose μ such that the overall rate of mutation is one. In this way, the length of the branch corresponds to the expected number of mutations per site along that branch (irrespective of the model)

More math-y: overall rate of mutation (expected number of non-redundant events in unit time $t=1$):

$$-\mu \cdot \text{trace}(\Pi \mathbf{Q})$$

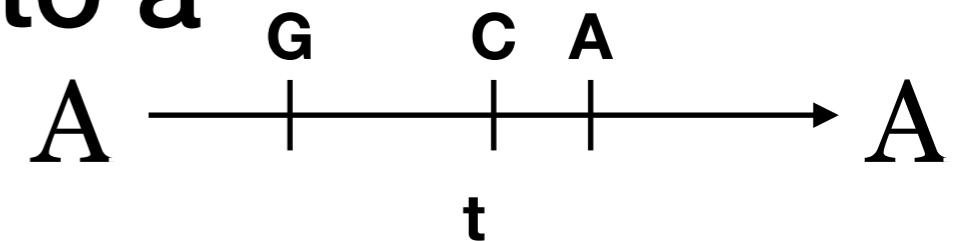
$$\mu = \frac{4}{3}$$

expected number of changes per site

$$P_{ij}(t) = 1/4 - 1/4e^{-4v/3}$$

$$P_{ii}(t) = 1/4 + 3/4e^{-4v/3}$$

We need a probability model for the evolution of an ancestral site into a descendant site



Homogeneous Continuous-time Markov chain

$$P(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{Qt}$$

Substitution model

Probabilities for the next event do not depend on the time point where the chain is

Event can happen at any time point (not just discrete steps)

Probabilities of the next state only depend on the current state

Ergodic and time reversible

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

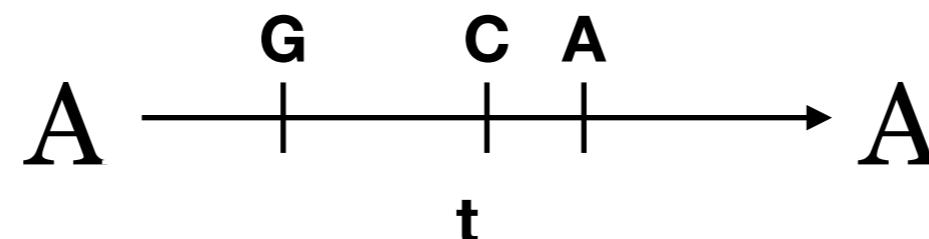
Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{\mathbf{Q}ut}$$

Branch lengths in expected number of substitutions, not time
We choose Q to select the substitution model we want

Substitution model

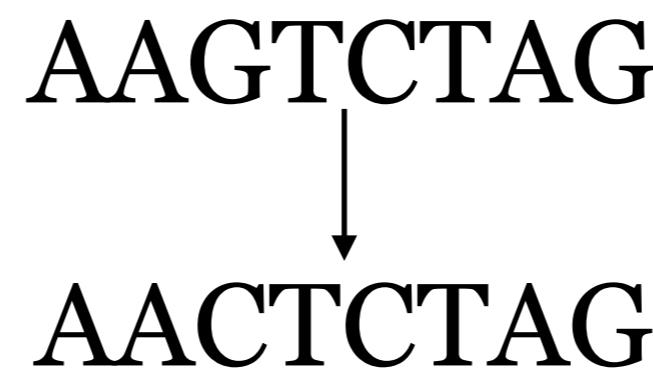
We need a probability model for the evolution of an ancestral site into a descendant site



$$P_{site}(t)$$

Assumption 2: We assume sites evolve independently

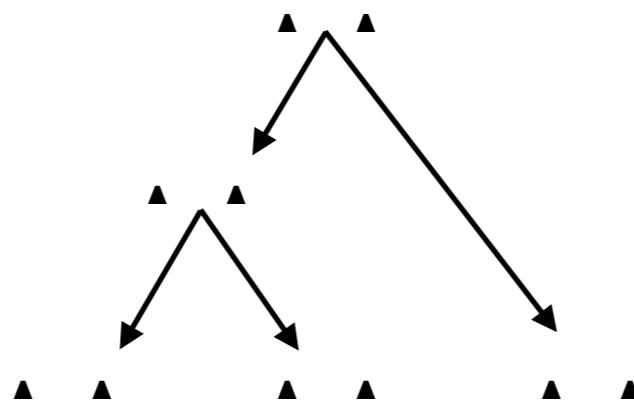
We need a probability model for the evolution of one ancestral sequence into a descendant sequence



Assumption 3: All sites evolve the same

$$P_{sequence}(t) = \prod P_{site}(t)$$

We need a probability model for the evolution of sequences along a phylogenetic tree



$$P_{tree} = \prod P_{branch}(t)$$

Assumption 1: The mutation process is the same at every branch of the tree

Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|c|}\hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline\end{array} = e^{\mathbf{Q}\mu t}$$

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Substitution model

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|c|}\hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline\end{array} = e^{\mathbf{Q}\mu t}$$

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Substitution model

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. FREE

N Galtier, M Gouy

Molecular Biology and Evolution, Volume 15, Issue 7, July 1998, Pages 871–879,
<https://doi.org/10.1093/oxfordjournals.molbev.a025991>

Published: 01 July 1998

Biologically?

Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|c|}\hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline\end{array} = e^{\mathbf{Q}ut}$$

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Substitution model

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. FREE

N Galtier, M Gouy

Molecular Biology and Evolution, Volume 15, Issue 7, July 1998, Pages 871–879,
<https://doi.org/10.1093/oxfordjournals.molbev.a025991>

Published: 01 July 1998



Regular article
Coevolving protein residues: maximum likelihood identification and relationship to structure ¹

David D. Pollock ¹ William R. Taylor ¹, Nick Goldman ²

Biologically?

Biologically?

Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|c|}\hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline\end{array} = e^{\mathbf{Q}ut}$$

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Substitution model

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. FREE

N Galtier, M Gouy

Molecular Biology and Evolution, Volume 15, Issue 7, July 1998, Pages 871–879,
<https://doi.org/10.1093/oxfordjournals.molbev.a025991>

Published: 01 July 1998



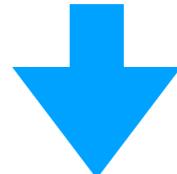
Regular article
Coevolving protein residues: maximum likelihood identification and relationship to structure ¹

David D. Pollock ¹ William R. Taylor ¹, Nick Goldman ²

Biologically?

Biologically?

Same rate for every site



Among-site rate variation (ASRV)

Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{\mathbf{Q}\mu t}$$

Substitution model

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Published: September 1994

Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods

Ziheng Yang

Journal of Molecular Evolution 39, 306–314(1994) | [Cite this article](#)

1861 Citations | 7 Altmetric | [Metrics](#)

Among-site rate variation (ASRV)

- Variation of evolutionary rates across sites modeled by a continuous distribution
- The rate of a specific site i is not constant but a random variable $r(i)$
- The likelihood for site i is calculated by integrating over all possible rates
- The distribution of rates is usually assumed to be Gamma
- Instead of the numerical integration, a discrete distribution for rates tends to be used

Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{\mathbf{Q}\mu t}$$

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Substitution model

Published: September 1994

Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods

Ziheng Yang

Journal of Molecular Evolution 39, 306–314(1994) | [Cite this article](#)

1861 Citations | 7 Altmetric | [Metrics](#)

Among-site rate variation (ASRV)

Site-specific rate variation

- Variation of evolutionary rates across sites modeled by a continuous distribution
- The rate of a specific site i is not constant but a random variable $r(i)$
- The likelihood for site i is calculated by integrating over all possible rates
- The distribution of rates is usually assumed to be Gamma
- Instead of the numerical integration, a discrete distribution for rates tends to be used

Homogeneous Continuous-time Markov chain

$$\mathbf{P}(t) = \begin{array}{|c|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{\mathbf{Q}\mu t}$$

Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Substitution model

Published: September 1994

Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods

Ziheng Yang

Journal of Molecular Evolution 39, 306–314(1994) | [Cite this article](#)

1861 Citations | 7 Altmetric | [Metrics](#)

A slow site can become fast

Site-specific rate variation

- Variation of evolutionary rates across sites modeled by a continuous distribution
- The rate of a specific site i is not constant but a random variable $r(i)$
- The likelihood for site i is calculated by integrating over all possible rates
- The distribution of rates is usually assumed to be Gamma
- Instead of the numerical integration, a discrete distribution for rates tends to be used

covarion - heterotachy

Published: March 1971

Rate of change of concomitantly variable codons

W. M. Fitch

Journal of Molecular Evolution 1, 84–96(1971) | [Cite this article](#)

93 Accesses | 84 Citations | [Metrics](#)

Comparative Study > *Math Biosci.* 1998 Jan 1;147(1):63–91.

doi: 10.1016/s0025-5564(97)00081-3.

Modeling the covarion hypothesis of nucleotide substitution

C Tuffley¹, M Steel

Affiliations + expand

PMID: 9401352 DOI: 10.1016/s0025-5564(97)00081-3



Summary

Homogeneous Continuous-time Markov chain

$$P(t) = \begin{array}{|c|c|c|c|} \hline & A & C & G & T \\ \hline A & & & & \\ \hline C & & & & \\ \hline G & & & & \\ \hline T & & & & \\ \hline \end{array} = e^{Qt\mu}$$

Substitution model

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

We choose if we want ASRV

Probabilities for the next event do not depend on the time point where the chain is

Event can happen at any time point (not just discrete steps)

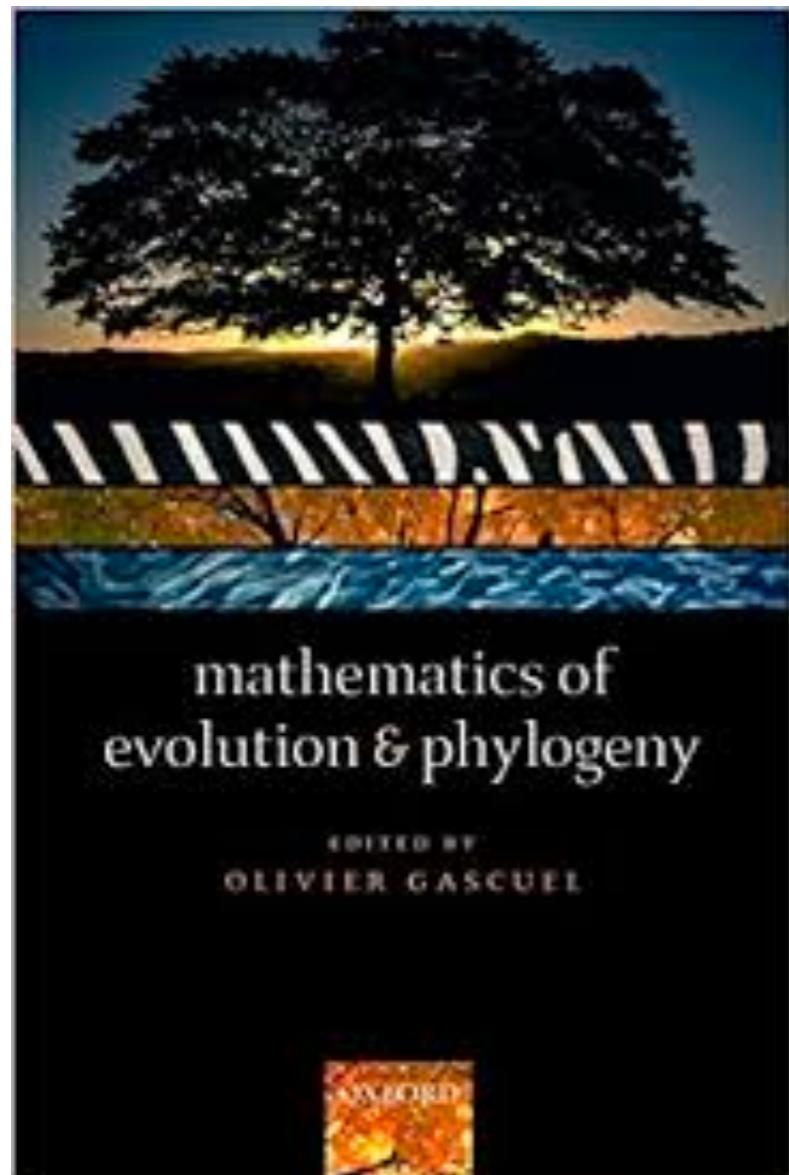
Probabilities of the next state only depend on the current state

Ergodic and time reversible

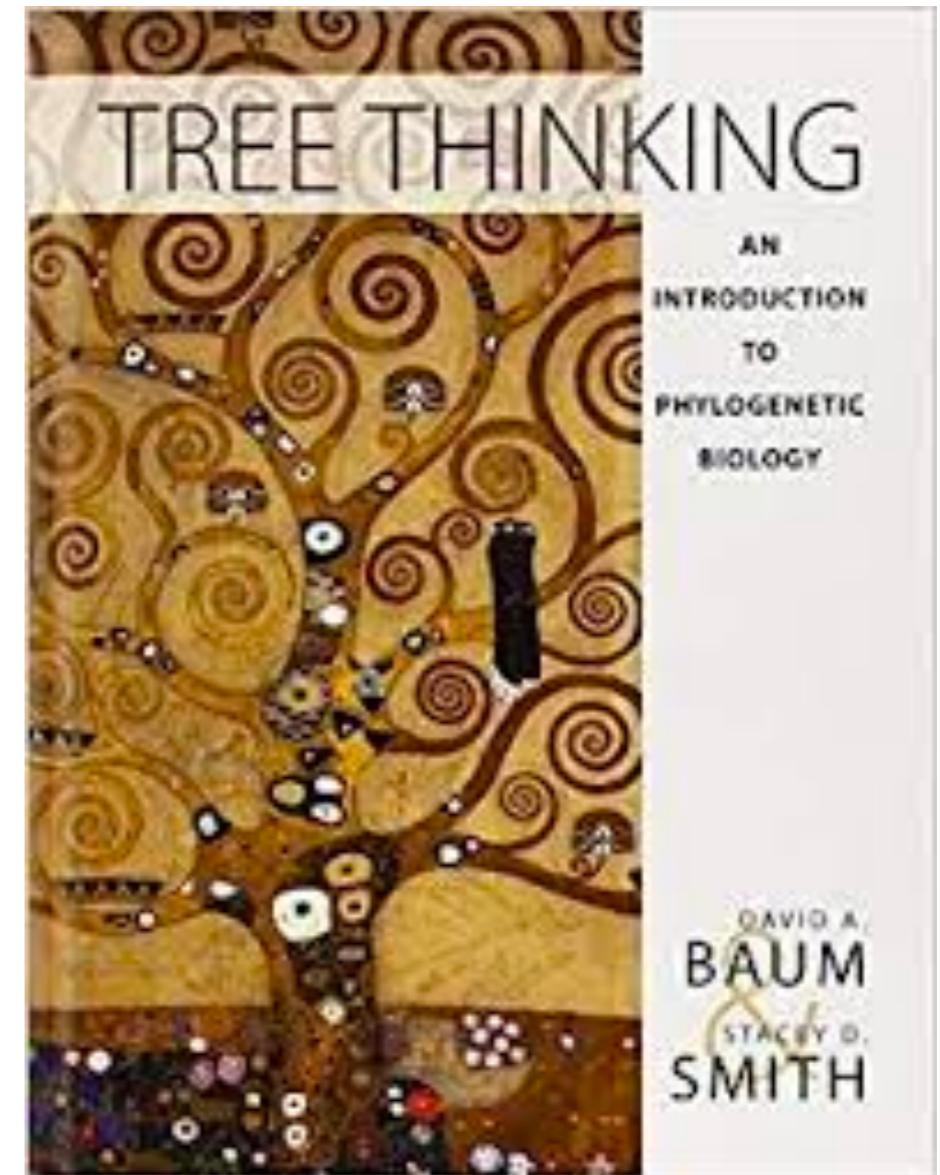
Branch lengths in expected number of substitutions, not time

We choose Q to select the substitution model we want

Further reading



Mathematics of evolution and phylogeny
Edited by Olivier Gascuel
Chapter 2



Tree thinking
David Baum and Stacey Smith
Chapter 8

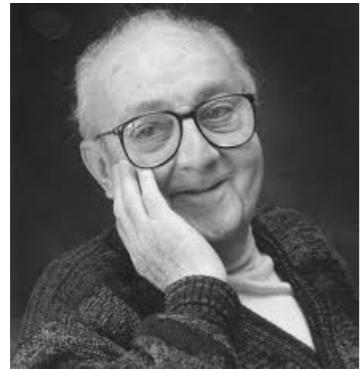
How can anything work under such unrealistic assumptions?

“The purpose of models is not to fit the data, but to sharpen the questions”

–Samuel Karlin, 11th Fisher Memorial Lecture

Explicit model is thought as a limitation of likelihood, but it is a strength

**All models are wrong,
but some models are
useful**



George Box
Founder of UW Stat
department

How useful?

It will depend on the
assumptions of the model
and the quality of the
input data

In-class activity

- **Time:** 5 minutes
- **Instructions:** Let's go over the different orthology methods that will be used in class

Model selection

Guest lecture recording by Rob Lanfear (IQ-Tree) in canvas:
botany563-lecture13.mpa

Important:

The video is not meant to be publicly available. You can watch it within canvas,
but should not download it or distribute it due to student confidentiality